A Method for Analysis of Expert Committee Decision-Making Applied to FDA Medical Device Panels

by

David André Broniatowski

S.B., Aeronautics and Astronautics, Massachusetts Institute of Technology, 2004
S.M., Aeronautics and Astronautics, Massachusetts Institute of Technology, 2006
S.M., Technology and Policy, Massachusetts Institute of Technology, 2006

SUBMITTED TO THE ENGINEERING SYSTEMS DIVISION IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY IN TECHNOLOGY, MANAGEMENT, AND POLICY
AT THE
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

JUNE 2010

Author_____
Engineering Systems Division
May 11, 2010

Certified by_____
Christopher L. Magee
Professor of the Practice of Mechanical Engineering and Engineering Systems
Thesis Supervisor

Certified by_____
Maria C. Yang
Assistant Professor of Mechanical Engineering and Engineering Systems
Committee Member

Certified by_____
Joseph F. Coughlin
Director, AgeLab
Committee Member

Accepted by_____
Nancy G. Leveson
Professor of Aeronautics and Astronautics, and Engineering Systems
Chair, Engineering Systems Division Education Committee

A Method for Analysis of Expert Committee Decision-Making Applied to FDA Medical Device Panels

by

David André Broniatowski

Submitted to the Engineering Systems Division on May 11, 2010 in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy in Technology, Management, and Policy

ABSTRACT

Committees of experts are critical for decision-making in engineering systems. This is because the complexity of these systems requires that information is pooled from across multiple specialties and domains of knowledge. The social elements of technical decision-making are not well understood, particularly among expert committees. This is largely due to a lack of methodology for directly studying such interactions in real-world situations. This thesis presents a method for the analysis of transcripts of expert committee meetings, with an eye towards understanding the process by which information is communicated in order to reach a decision. In particular, we focus on medical device advisory panels in the US Food and Drug Administration. The method is based upon natural language processing tools, and is designed to extract social networks in the form of directed graphs from the meeting transcripts which are representative of the flow of information and communication on the panel. Application of this method to a set of 37 meetings from the FDA's Circulatory Systems Devices Panel shows the presence of numerous effects. Prominent among these is the propensity for panel members from similar medical specialties to use similar language. Furthermore, panel members who use similar language tend to vote similarly. We find that these propensities are correlated – i.e., as panel members' language converges by medical specialty, panel members' votes also converge. This suggests that voting behavior is mediated by membership in a medical specialty and supports the notion that voting outcome is, to some extent, dependent on an interpretation of the data associated with training, particularly when a small number of interpretations of the data are possible. Furthermore, there is some preliminary evidence to suggest that as clinical trial data ambiguity and difficulty of decision-making increases, the strength of the mediating effect of medical specialty decreases. Assuming a common decision is reached, this might indicate that committee members are able to overcome their specialty perspective as the committee jointly deals with hard problems over longer periods of time. In cases where the panel's vote is split, a lack of linguistic coherence among members of the same medical specialty correlates with a lack of linguistic coherence among members who vote the same way. This could be due to the presence of multiple interpretations of the data, leading to idiosyncratic or value-based choice. We also find that voting outcome is associated with the order in which panel members ask questions – a sequence set by the committee chair. Members in the voting minority are more likely to ask questions later than are members in the voting majority. Voting minority members are also more likely to be graph sinks (i.e., nodes in a social network that have no outflow) than are voting majority members. This suggests an influence mechanism on these panels that might be associated with framing – i.e., later speakers seem to be less able to convince other panel members to

discuss their topics of interest contributing to these members' minority status. These results may have some relation to FDA panel procedures and structure. Finally, we present a computational model that embodies a theory of panel voting procedures. Model results are compared to empirical results and implications are drawn for the design of expert committees and their associated procedures in engineering systems.

Thesis Supervisor: Christopher L. Magee
Title: Professor of the Practice of Mechanical Engineering and Engineering Systems

ACKNOWLEDGMENTS

An undertaking such as a doctoral dissertation must always require more thanks to more people than can be expressed in words. Nevertheless, one must try. Throughout my time at MIT, I have had a number of mentors and friends who have guided me through difficult times. I would primarily like to thank the members of my doctoral committee.

My committee chair, **Prof. Chris Magee**, took me in as a student at a time when I must have seemed a brash, unfocused and certainly unfunded space cadet. Prof. Magee believed in my potential and allowed me the right combination of guidance and autonomy required to bring this work to fruition. Even more important is his guidance, both in matters of research and in matters of life, exemplified by our discussions that range in topic from career advice to ethics. Prof. Magee displays a combination of three qualities rarely found in academia -- a practical bent combined with deep theoretical understanding and an abiding respect for the humanistic consequences of academic research. The sometimes harsh academic environment can often lead individuals to categorize their colleagues according to tightly-defined academic and disciplinary boundaries. The concomitant biases imposed by this rigid thinking can be mildly dehumanizing. Working with Prof. Magee has been a refreshing reminder of the depth of the humanity that resides in all of us.

Prof. Maria Yang has always been an enthusiastic supporter of the research direction embodied in this work. In addition to her technical excellence, Prof. Yang's empathy, kindness, and balanced approach to life, has much to recommend to a young PhD student who is trying to find his way through the academic jungle. In an institution where everyone is routinely busy, Prof. Yang makes time for others. It is precisely this sort of selflessness that is needed if we, as a society, are to successfully tackle the large-scale technical challenges awaiting us in the future -- no one can do it alone. Prof. Yang is therefore a leader in the best sense of the word -- one who acts with humility and perhaps even without explicit intention to influence others. She inspires collaborators; not followers, and in so doing, empowers those around her.

Dr. Joe Coughlin is perhaps one of the "hidden gems" of the Engineering Systems Division. In addition to the tremendous work that is done in the AgeLab, Dr. Coughlin possesses a formidable command of theory that is directly relevant to the emerging field of Engineering Systems; especially as it relates to the social aspect. Among faculty that largely comes from an engineering

background, Dr. Coughlin provides a voice that is tempered by experience in the policy realm. He therefore has the ability to temper his theoretical understanding with practical experience, enabling deep insight into problems of social importance. It has been my privilege to discuss these insights with him. Although MIT has, at times, been called a "praise-free zone", Dr. Coughlin has always been kind and supportive, and possesses the one quality that is absolutely necessary for survival in any PhD program -- a sublime sense of humor.

A number of individuals have been extremely influential in my intellectual growth over the past four years:

It was **Prof. Joel Moses** who first introduced me to the ideas of expertise as a phenomenon worth studying in its own right, as well as the notion of (professional and national) culture influencing patterns of thought. When I suggested to him that this relationship might extend to language, Prof. Moses pointed me in the direction of the Sapir-Whorf hypothesis. Our discussions, although too infrequent, have always prompted me to examine my assumptions and to immerse myself in literatures from other cultures and other periods of time. I have developed a deep interest in Idealist philosophy due to his influence, and I feel that my life has been enriched as a result. Perhaps Prof. Moses' greatest lesson is what I have yet to learn: indeed, attempts to simply understand his perspective have led me down the path of my own self-knowledge – for this I am grateful. What remains to be said on this topic is ineffable. I can only hope that this work might inspire in him some small amount of nakhes.

Prof. Whitman Richards has provided me with invaluable guidance and advice, especially as regards my own academic growth. His course on Cognitive Architectures blew open the doors of my mind, and fundamentally changed the way that I think about social and cognitive systems. It was due to his influence that I was able to take the crucial step of generating social networks for my analysis. In addition, he has inspired me to explore academic pathways that I didn't even know existed. I hope that we will continue to be able to work together in the future.

Outside of MIT, **Rabbi Dov Berish Ganz** has provided me with an opportunity both to sharpen my intellectual skills and to develop my ethical awareness through the study of Jewish texts. I would like to think that this work represents some of his influence in the form of "computational mussar". Although MIT can be an environment in which kindness is in short supply, Rabbi Ganz, and his wife, **Rivky Ganz**, possess kindness, wisdom, and hospitality in abundance. The city of Cambridge is immeasurably enriched by their presence. Thanks, as well, to **Rabbi Shlomo Yaffe** for our time spent studying chassidut. His intelligence,

breadth of knowledge, and continued interest in this project have inspired me. I hope that this work may perhaps represent the beginnings of a droplet in the coming fountain of wisdom.

Special thanks must be proffered to the MIT-Portugal Program, for financial support throughout the last three years. Thanks, especially, to **Terra Cholfin**, for her constant help in dealing with the MIT bureaucracy and enabling me to take my research so frequently on the road. Thanks, as well, to **Prof. Annalisa Weigel**, for her support of my first semester in ESD and for her encouragement in my pursuit of an interdisciplinary background.

No one can get through a PhD program alone. One learns the most from ones friends. In particular, thanks to my close friends and MIT colleagues **Chris Lawson**, **Nichole Argo**, **Matt Richards** and **Nirav Shah** for interesting discussions and shared experiences. Thanks, especially, to **Jason Bartolomei** who provided financial and moral support when it appeared that there were no other options. Outside of MIT, **Ariel Evan Mayse**, **Chava Evans**, **Mark Dredze**, and **Zev Berger** have all contributed greatly to my intellectual and personal growth. Someone once told me that our words constantly create our reality. Words between friends make this a reality worthy of that creation. Thanks, as well, to the **Systems Engineering Advancement Research Initiative** for providing me with office space and an "academic home" within ESD despite my sometimes sticking out like a sore thumb.

I owe my biggest thanks to the unending support of my family, and it is to them that this thesis is dedicated. My brother, **Daniel Broniatowski**, and his wife, **Holly Kilim**, have never ceased in their hospitality toward me. Holly's father, **Rafi Kilim**, helped me to keep things in perspective before my general exams. My uncle, **Warren Grundfest**, first exposed me to the idea that health care may be considered a complex engineering system and supported me in my decision to follow this path. My grandparents, **Jack and Karolyn Grundfest**, have taught me so much about maintaining a graceful calm in the face of uncertainty – I can only hope to emulate their example and to partake of a fraction of their wisdom. My grandmother, **Irmgard Broniatowski**, has taught me by example not to lose hope in the darkest times, and the memory of my grandfather and namesake, **Chaim André Broniatowski**, continues to inspire me in all that I do. Finally, and most importantly, I would like to thank my parents, **Michael and Sharon Broniatowski**. Growing up, our dinner table conversations have made me aware of the different professional cultures within medicine. Much more importantly, I would like to thank them for their love and support, and especially for putting up with my constant kvetching.

To all these people, and to those who have unintentionally not been named, I express my gratitude. You are all responsible for the results presented in this dissertation. All errors are, of course, mine alone.

LIST OF FIGURES

Figure 1: Medical devices are classified into three categories based upon risk to the patient. PMA = Pre-Market Approval; GMP = Good Manufacturing Practices. 46
Figure 2: Medical devices are classified into three categories based upon risk to the patient. Diagram sourced from (Maisel 2004). 47
Figure 3: Standard Layout for FDA CDRH Advisory Panel Meeting. 49
Figure 4: Coherence of group centroid with respect to final centroid of voting members. The horizontal axis, representing the utterance number in the discourse, represents progress through the discourse. The vertical axis is coherence as measured with respect to the final position of the voting members. Each curve corresponds to a different group present at the meeting (Non-voting panel members, FDA representatives, sponsors, and voting members) 66
Figure 5: Breakdown of sponsor's presentation by speaker/section. Each curve corresponds to a different speaker, each of whom presented a different phase of the sponsor's talks (introduction, clinical data, statistical method, and physician endorsement) 67
Figure 6: A plate-notation representation of the Latent Dirichlet Analysis algorithm (Blei, Ng, et al. 2003) 73
Figure 7: A plate notation representation of the Author-Topic model from (Rosen-Zvi et al. 2004). Authors are represented by a multinomial distribution over topics, which are in turn represented by a multinomial distribution over all words in the corpus. 80
Figure 8: A comparison of perplexity values for three different hyperparameter conditions tested. Fitted priors generally have slightly lower perplexity, particularly for longer meetings. 84
Figure 9: Sample output from the Author-Topic model run on the FDA Circulatory Systems Devices Advisory Panel Meeting for March 4th, 2002. This chart is the per-speaker topic distribution for one of the panel members. 88
Figure 10: A graph of the meeting of the FDA Circulatory Systems Devices Advisory Panel Meeting held on March 5, 2002. Node size is proportional to the number of words spoken by the corresponding speaker. Random seed = 613. Graphs were generated using UCINET. 95
Figure 11: Social network of the FDA Circulatory Systems Devices Advisory Panel Meeting held on March 5, 2002. Threshold value is determined using the binomial test described above. Node size is proportional to the number of words spoken by the corresponding speaker. Random seed =

201.657. 2100th draw from MCMC algorithm. Graphs were generated using UCINET. This iteration shows the presence of two separate discussion groups. Note that voting members 5 and 6, both bridging members in Figure 10, are now disconnected. This is due to their small number of words contributed. 97
Figure 12: A second iteration of the meeting of the FDA Circulatory Systems Devices Advisory Panel Meeting held on March 5, 2002. Threshold value is determined using the binomial test described above. Node size is proportional to the number of words spoken by the corresponding speaker. Random seed = 201.657. 2200th draw from MCMC algorithm. Graphs were generated using UCINET....98
Figure 13: Average of 200 graphs for the meeting of the FDA Circulatory Systems Devices Advisory Panel Meeting held on March 5, 2002. A heavy line indicates a strong link (linked in >100 graphs). A light line indicates that the speakers are linked more frequently than the global average of all speakers. Spurious links have been eliminated. 100
Figure 14: Average of 200 iterations for the meeting of the FDA Circulatory Systems Devices Advisory Panel Meeting held on March 5, 2002. Iterations use a binomial threshold value for each of ten topics. Heavier lines indicate stronger links (linked in >100 iterations), whereas lighter lines indicate weaker links (> than the global average). All links shown are stronger than the global average of all speakers. Remaining links have been deleted. 101
Figure 15: Average of 200 iterations for the meeting of the FDA Circulatory Systems Devices Advisory Panel Meeting held on January 13, 2005. Iterations use a constant threshold value for each of ten topics. A heavy line indicates a strong link (linked in >100 iterations). A light line indicates that the speakers are linked more than the global average of all speakers. Remaining links have been deleted. 103
Figure 16: Average of 200 iterations for the meeting of the FDA Circulatory Systems Devices Advisory Panel Meeting held on January 13, 2005. Iterations use a binomial threshold value for each of ten topics. Heavier lines indicate stronger links, whereas lighter lines indicate weaker links. All links shown are stronger than the global average of all speakers. Remaining links have been deleted....104
Figure 17: Perplexity vs. number of topics for the meeting of the FDA Circulatory Systems Devices Panel held on July 9, 2001. T, the number of topics, is equal to 28, using the procedure described above. Horizontal lines indicate the 5th and 95th percentiles for perplexity for a 27 topic model fit....107
Figure 18: *A priori* probability distribution for links between speakers in the April 21, 2004 meeting with 28 topics. The median of this distribution is

0.0356; whereas 1/28 = 0.0357. The assumption of a symmetric Dirchlet prior distribution implies that this distribution holds for all speakers until it is updated with data observed from the transcripts. 109
Figure 19: Sample histogram of linkage frequency for an FDA Advisory Panel meeting of April 21, 2004. The horizontal axis is the link weight (i.e., the frequency with which author-pairs are connected over 200 samples from the AT model). The vertical axis is the link frequency of links with the weight specified by the abcissa (i.e., the number of author-pairs that are connected with the frequency specified by the abcissa). Note the existence of two modes located at the extremes of the distribution. 111
Figure 20: Graph of the FDA Circulatory Systems Advisory Panel meeting held on December 5, 2000. This meeting yielded a consensus approval of the medical device under analysis. Node shape represents medical specialty. The committee chair is in black. 113
Figure 21: Graph of the FDA Circulatory Systems Advisory Panel meeting held on October 27, 1998. This meeting yielded an approval of the medical device under analysis, with only one dissenter (in red). Node shape represents medical specialty. The committee chair is labeled and did not vote. The voter in black was not present for the vote. 114
Figure 22. Graph of the FDA Circulatory Systems Advisory Panel meeting held on April 21, 2004. This meeting yielded an approval of the medical device under analysis, although the panel was split (blue, in favor; red against). Node shape represents medical specialty. The committee chair is in black. 115
Figure 23: Graph of the FDA Circulatory Systems Advisory Panel meeting held on June 6, 1998. This device was not approved. Node shape represents medical specialty. The committee chair is in black. Non-approval votes are in red; approval votes are in blue. In this meeting, vote is not correlated with medical specialty. 116
Figure 24: Graph of the FDA Circulatory Systems Advisory Panel meeting held on June 23, 2005. Node color represents the vote (red is against humanitarian device exemption, blue is in favor of humanitarian device exemption, black is abstention. The committee chair is also black. Node shape represents medical specialty. 118
Figure 25: Comparison of prior and posterior distribution of link probabilities for two strongly-linked voting members during the April 21, 2004 meeting. An ideal observer would place the link probability threshold around 0.04, indicating that a joint probability greater than this value would signal a link with very high likelihood. 119
Figure 26: Weighted graph representation of the meeting held on March 5, 2002. Link weights reflect the likelihood that a given edge is due to sharing a topic compared to a background prior distribution. Note that this graph

has a similar connectivity pattern to that shown in Figure 13, although it is somewhat denser due to low-likelihood links (e.g., those near 1) 121
Figure 27: Weighted graph representation of the meeting held on March 5, 2002. Link weights reflect the likelihood that a given edge is due to sharing a topic compared to a background prior distribution. Note that this graph has a similar connectivity pattern to that shown in Figure 15, although it is somewhat denser due to low-likelihood links (e.g., those near 1). 122
Figure 28: First segment of the January 13, 2005 Circulatory Systems Devices Panel Meeting. At this point in the meeting, voting members had not yet expressed any preferences regarding voting. Rather, committee members were listening to the open public hearing and sponsor presentations. Data include utterances 1-377 of 1671 total utterances.....124
Figure 29: Second segment of the January 13, 2005 Circulatory Systems Devices Panel Meeting. This graph shows that, at this point in the meeting, Voting Members 5, 7, 8, 10, 11 and 12 had begun discussing the statistical elements of the clinical trial design. Five of the six surgeons present have not yet expressed utterances. Data include utterances 378-589 of 1671 total utterances.....125
Figure 30: Third, and final, segment of the January 13, 2005 Circulatory Systems Device Panel Meeting. This graph shows that, after lunch, the surgeons in the room, who were previous silent, seemed to align in favor of device approval. Voting Members 8, 9, 10 and 12 seemed to maintain their relative positions between the second and third segments. Data include utterances 590-1671.....126
Figure 31: Before-lunch segment of the March 5th, 2002 Circulatory Systems Devices Panel Meeting. This graph shows that, at this point in the meeting, voting members had largely aligned themselves into blocs that would later vote similarly. Data include utterances 1-703 of 1250 total utterances.....127
Figure 32: After-lunch segment of the March 5th, 2002 Circulatory Systems Devices Panel Meeting. This graph shows that, by the second half of the meeting, those who would later vote against device approval had become more strongly linked to those who would later support device approval. This pattern perhaps reflects attempts by the approval voters to convince the non-approval voters to vote differently. Data include utterances 704-1250 of 1250 total utterances. 128
Figure 33: Before-lunch segment of the April 21st, 2004 Circulatory Systems Devices Panel Meeting. This graph shows well-defined coalitions having been formed relatively early in the meeting. It is interesting that voting patterns seem to largely respect the boundaries of particular medical specialties (i.e., surgeons vs. cardiologists). Data include utterances 399-876 of 1822 total utterances.....130

Figure 34: After-lunch segment of the April 21st, 2004 Circulatory Systems Devices Panel Meeting. This graph shows that the well-defined coalitions of the before-lunch segment have broken down – particularly the anti-device coalition. This may well be due to attemtps by members of one coalition to influence the other, leading to cross-coalition dialogue.. Data include utterances 877-1822 of 1822 total utterances. 131
Figure 35: Time series for two speakers on topic #13 during the meeting held on January 13, 2005. 134
Figure 36: Cross-correlation of the two time series shown in Figure 35. 135
Figure 37: A cross-correlation function with two peaks, representing two speakers who are equally involved in leading conversation on this topic. 136
Figure 38: Edge direction distribution for two speakers, one of who clearly leads the other. Both speakers were voting members in the meeting held on January 13, 2005. 138
Figure 39: Edge direction distribution for two speakers, one of whom clearly lags the other. Both speakers were voting members in the meeting held on January 13, 2005. 139
Figure 40: Edge direction distribution for two speakers, neither of whom clearly lags the other. Both speakers were voting members in the meeting held on January 13, 2005. 140
Figure 41: Directed network representation of the FDA Circulatory Systems Advisory Panel meeting held on January 13, 2005. Node size increases with the number of words spoken by that author; node shape represents medical specialty. Non-approval votes are red; approval votes are blue; non-voters are black. Each speaker's top five words are listed, as is each edge's link frequency. This diagram is generated using the dot algorithm (Gansner and North 1999) 141
Figure 42: Directed network representation of the FDA Circulatory Systems Advisory Panel meeting held on July 9, 2001. Node size increases with the number of words spoken by that author; node shape represents medical specialty. Non-approval votes are red; approval votes are blue; non-voters are black. This diagram is generated using the dot algorithm (Gansner and North 1999) 142
Figure 43: A Kruskal-Wallis test shows no significant difference between the air-time proportions of majority and minority voters ($p=0.86$) for the 17 meetings in which a split-vote existed. 149
Figure 44: Box plots for the four most strongly-represented specialties. Note that more "clinical" specialties (surgeons and electrophysiologists) tend to speak less than the more "medical" specialties (cardiologists and statisticians). 153
Figure 45: Histogram of Specialty Cohesion Percentiles for the 37 meetings in our sample. The empirical specialty cohesion percentile distribution's

cumulative distribution function is significantly less than that of the background distribution (one-sided Kolmogorov-Smirnov test; p=0.0045) indicating that the empirical distribution has more probability density concentrated near unity and away from zero. 155
Figure 46: Cumulative Distribution Plot of Specialty Cohesion Percentiles for the 37 meetings in our sample. The empirical specialty cohesion percentile distribution's cumulative distribution function is significantly less than that of the background distribution (one-sided Kolmogorov-Smirnov test; p=0.0045) indicating that the empirical distribution has more probability density concentrated near unity and away from zero. 156
Figure 47: Histogram of Vote Cohesion Percentiles for the 11 meetings with a minority of size 2 or greater. The empirical vote cohesion percentile distribution's cumulative distribution function is significantly less than that of the background distribution (one-sided Kolmogorov-Smirnov test; p=0.015) indicating that the empirical distribution has more probability density concentrated near unity and away from zero. 158
Figure 48: Cumulative Distribution Plot of Specialty Cohesion Percentiles for the 11 meetings with a minority with two or more voting members. The empirical vote cohesion percentile distribution's cumulative distribution function is significantly less than that of the background distribution (one-sided Kolmogorov-Smirnov test; p=0.015) indicating that the empirical distribution has more probability density concentrated near unity and away from zero. 159
Figure 49: Scatter plot of Vote Cohesion percentile vs. Specialty Cohesion percentile for 11 meetings in which there was a minority of two or more. Vote and specialty cohesion percentiles are positively associated (Spearman Rho =0.79; p=0.0061). Each datapoint is labeled by its corresponding meeting ID, as catalogued in Appendix 3. Datapoints are also color-coded by the proportional size of the minority in each meeting, suggesting that this effect holds independent of proportional minority size. 160
Figure 50: Kruskal-Wallis non-parametric ANOVA finds a significant difference between the median speaking order rank of voting majority and voting minority voting members in the 17 meetings in which there was a voting minority (abstentions were not included); p=0.0008. Voting minority members speak later than majority members do. 162
Figure 51: Kruskal-Wallis non-parametric ANOVA finds a significant difference between the median speaking order rank of voting majority and voting minority voting members in the 11 meetings in which there was a voting minority with two or more voting members (abstentions were not included); p=0.011. Voting minority members speak later than voting majority members do. 163

Figure 52: Kruskal-Wallis non-parametric ANOVA finds a significant difference between the outdegree of voting majority and voting minority panel members in the 17 meetings in which there was a majority (abstentions were not included); p=0.045. There is no observable effect for indegree (p=0.67) or undirected degree (p=0.37). 164
Figure 53: Kruskal-Wallis non-parametric ANOVA finds a significant difference between the outdegree of voting majority and voting minority panel voting members in the 11 meetings in which there was a majority of size two or larger (abstentions were not included); p=0.058 165
Figure 54: Members of the voting minority (in favor of device approval) speak significantly later than do members of the voting majority (against device approval) in the 10 meetings in which the panel voted not to approve the devices (p=0.0025) 169
Figure 55: Members of the voting minority (against device approval) do not speak significantly later than do members of the voting majority (in favor of device approval) in the 7 meetings in which the panel voted not to approve the devices (p=0.12). By inspection, there is a non-significant trend for the voting minority to speak later than does the voting majority. 170
Figure 56: Members of the voting minority (in favor device approval) do not have significantly smaller outdegrees than do members of the voting majority (against device approval) in the 10 meetings in which the panel voted not to approve the devices (p=0.27). 171
Figure 57: Members of the voting minority (against device approval) have marginally significantly smaller outdegrees than do members of the voting majority (in favor of device approval) in the 7 meetings in which the panel voted to approve the devices (p=0.056). 172
Figure 58: Kruskal-Wallis non-parametric ANOVA finds a significant difference between proportional voting minority size in the 35 meetings in which there was a voting minority and at least one lead reviewer in the voting minority (p=0.0006). Similar results are obtained when focusing on the subset of 17 meetings with a voting minority (p=0.027). There is insufficient data to obtain a similar result for the subset of 11 meetings with a voting minority of 2 or more (p=0.33), although the direction of the trend remains the same. 179
Figure 59: Maximum normalized outdegree is significantly associated with meeting length (Spearman rho=0.50; p=0.04). Datapoints are labeled by the meeting ID assigned in Appendix 3. There is no significant association between location of first minority member in the speaking order and meeting length (p=0.50). 181
Figure 60: Maximum normalized outdegree is significantly associated with voting minority proportional size (Spearman rho=0.62; p=0.0082) for the 17

meetings in which there is a minority. Datapoints are labeled by the meeting ID assigned in Appendix 3. 182
Figure 61: Plot of Meeting Length vs. voting minority proportional size. Meeting length is significantly positively associated with voting minority proportional size (Spearman Rho = 0.53; $p=7.1 \times 10^{-4}$). Decisions that are likely to have been clear or ambiguous are labeled. 183
Figure 62: Directed Graph representation of meeting held on June 23, 2005. Luo's hierarchy metric = 0.35. 184
Figure 63: Directed Graph representation of meeting held on June 23, 2005, with the committee chair included. Luo's hierarchy metric = 0.78. 185
Figure 64: Directed Graph representation of meeting held on October 27, 1998. Luo's hierarchy metric = 0. 186
Figure 65: Directed Graph representation of meeting held on October 27, 1998, with the committee chair included. Luo's hierarchy metric = 0. 188
Figure 66: Distribution of chair impacts for the set of 37 meetings analyzed. This distribution shows a bimodal structure. 189
Figure 67: The impact of the committee chair seems to be associated with meeting date. The vertical axis represents the number of days since January 1st, 1900. 190
Figure 68: Impact of chair vs. meeting date for each of the 17 meetings in which there was a voting minority. Note that after March 4, 2002, chair impact seems to increase for most meetings. Each meeting is labeled by its corresponding ID. 191
Figure 69: Half-day meetings were not held after March 4, 2002. These later meetings are marked in red, whereas earlier meetings are in blue. Each meeting is labeled by its corresponding ID. 192
Figure 70: Schematic of the model outlined in this chapter. 199
Figure 71: Plot of Simulated Specialty Cohesion vs. Simulated Vote Cohesion. Spearman Rho = 0.57; $p=8.97 \times 10^{-16}$. Proportional minority size (including abstentions) is represented in color. 207
Figure 72: Members of the simulated voting minority tend to speak later than do members of the simulated voting majority ($p<0.0001$). 208
Figure 73: In the absence of a pre-set speaking order, members of the simulated voting minority do not tend to speak later than do members of the simulated voting majority ($p=0.69$). 213
Figure 74: Analysis of Covariance Plot showing the effect of ambiguity on correlation between Specialty Cohesion Percentile and Vote Cohesion Percentile 218
Figure 75: Analysis of Covariance Plot showing the effect of complexity on correlation between Specialty Cohesion Percentile and Vote Cohesion Percentile 219

Figure 76: Three interacting orders or layers. Each one constrains the layer immediately above it. Thus very clear data would constrain the set of possible interpretations, etc.....253

LIST OF TABLES

Table 1: Listing of the top five log-entropy weighted words for each speaker. 63
Table 2: "Confusion Matrix" for stakeholder cluster analysis. ($p = 9.97 \times 10^{-5}$; $\chi^2 = 15.14$; df = 1) 64
Table 3: The top five word-stems for one run of the AT model on the corpus for the Circulatory Systems Devices Panel Meeting of March 4, 2002. 88
Table 4: Results of the Author-Topic Model applied to a transcript of the Circulatory Systems Devices Panel Meeting of Nov. 20, 2003. Each row of this table corresponds to a different voting member. Topics correspond to conditions of approval for the final vote. 91
Table 5: Different types of speakers identified by the AT model. A frequent, focused speaker tends to drive topic formation, whereas a rare speaker tends to be assigned to topics defined by others. Multi-focus, or interdisciplinary, speakers may serve as mediators. These sample results have been generated from actual panel meetings. 93
Table 6: 4-way ANOVA showing the effects of Gender, Medical Specialty, h-Index, and Age on air-time for our sample of 37 meetings. In this analysis, air-time has been normalized and a logit transform has been applied to enable comparisons across meetings. When race is included as an explanatory variable, it fails to reach significance ($p=0.20$), suggesting no identifiable effect of race. Medical Specialty captures most of the variance in air-time, followed by h-Index, gender and age. 146
Table 7: 4-way ANOVA showing the effects of Gender, Medical Specialty, h-Index, and Age on air-time for the subset of 17 meetings in which there was a minority. In this analysis, air-time has been normalized and a logit transform has been applied to enable comparisons across meetings. Here, most of the variance is captured by h-Index followed by Medical Specialty and age. Gender fails to reach significance as an explanatory variable. 147
Table 8: 4-way ANOVA showing the effects of Gender, Medical Specialty, h-Index and Age on voting outcome for the 17 meetings in which there was a voting minority. In this analysis, voting outcome is a dichotomous variable, thereby violating the ANOVA assumptions. Only gender has a significant effect on voting outcome. 150
Table 9: A chi-square test examining the impact of gender on voting outcome for the 17 meetings in which a minority existed shows a significant result ($\chi^2=8.29$;dof=1;$p=0.0040$) with women more likely to be in the majority. 151
Table 10: There is no significant relation between medical specialty and voting behavior ($\chi^2=4.29$;dof=8;$p=0.83$) 152

Table 11: 2-way ANOVA Table showing effect of Outdegree and Speaking Order on vote (majority vs. minority) for those 17 meetings in which there is a minority. Although the ANOVA assumptions are not met, an effect of Speaking Order is still evident (cf. Lunney 1970). The absence of an effect due to outdegree suggests that the variance in speaking order accounts for the variance in voting behavior as well as in outdegree. 166
Table 12: 2-way ANOVA Table showing effect of Outdegree and Speaking Order on vote (voting majority vs. voting minority) for those 11 meetings in which there is a minority with at least two members. Although the ANOVA assumptions are not met, an effect of Speaking Order is still evident. The absence of an effect due to Outdegree suggests that the variance in speaking order accounts for the variance in voting behavior as well as in outdegree....167
Table 13: Analysis of the 17 meetings with a voting minority indicates that members of the minority are more likely to be graph sinks than are members of the majority (χ^2 = 4.92; dof=1; p=0.026)....173
Table 14: Analysis of the 11 meetings with a voting minority including at least two members indicates that members of the voting minority are more likely to be graph sinks than are members of the voting majority (χ^2 = 4.66; dof=1; p=0.031). 173
Table 15: Analysis of the 17 meetings with a voting minority shows that members of the voting minority are more likely to the last speaker than are members of the voting majority (χ^2 =5.22; dof=1; p=0.022) 174
Table 16: Analysis of the 11 meetings with a voting minority of size two or more shows that members of this voting minority are not more likely to the last speaker than are members of the voting majority (χ^2 =0.94; dof=1; p=0.33)....175
Table 17: Analysis of the 6 meetings with a voting minority of size one only shows that members of the voting minority are more likely to the last speaker than are members of the voting majority (χ^2 =12.36; dof=1; p=0.00044). Of the three voting minority members who are the last speaker, two are graph sinks and one is a graph isolate (outdegree and indegree are both 0). 175
Table 18: Table of Precision, Recall, and F-Score for the data shown above. The graph sink method has a consistently higher precision and F-score, and is lower on recall only in the case of 17 meetings. 177
Table 19: Model Input Variables....197
Table 20: Kolmogorov-Smirnov tests show no significant differences between the empirical and simulated distributions for vote cohesion and for specialty cohesion. 209

Table 21: A chi-square test shows that the simulated data distribution of panel meeting outcomes does not match the empirical distribution (χ^2 =20.00; dof=2; p=4.54 x 10^{-5}) 210
Table 22: A chi-square test shows that the simulated data distribution of panel meeting outcomes does not match the empirical distribution (χ^2 =3.02; dof=2; p=0.22) 210
Table 23: Kolmogorov-Smirnov tests show no significant differences between the empirical and simulated distributions for specialty and vote cohesion or for their percentiles. 211
Table 24: Kolmogorov-Smirnov tests show significant differences between the empirical and simulated distributions for specialty and vote cohesion percentiles. 214
Table 25: Kolmogorov-Smirnov tests shows significant differences between the empirical and simulated distributions for vote cohesion and specialty and vote cohesion percentiles. 215
Table 26: 6-way Analysis of Variance showing the impact of Complexity, Mean Breadth, Mean Depth, Depth Dispersion, Openness and Ambiguity on Specialty Cohesion Percentile 216
Table 27: 2-way Analysis of Variance showing the impact of Complexity and Ambiguity on Vote Cohesion Percentile 217
Table 28: 5-way Analysis of Variance showing the impact of Diversity, Complexity, Mean Breadth, Mean Breadth, Quality and Ambiguity on proportional minority size. 220
Table 29: 4-way Analysis of Variance showing the impact of Mean Breadth, Mean Depth, Quality and Ambiguity on correct vote outcome. Although correct vote outcome is a dichotomous variable, the analysis is still qualitatively instructive (Lunney 1970). 221
Table 30: 4-way ANOVA showing the effects of Age, Medical Specialty, Gender, and Race on h-index for our sample of 37 meetings. All variables reach significance. 268

C h a p t e r 1

INTRODUCTION

<<Diodore de Sicile [marg: lib.2. Biblioth. Hist.] explique l'invention des Langues de cette maniere. Les hommes faisant leurs premiers coups d'essai pour parler, prononcerent d'abord des sons qui ne signifioient rien: puis, aprés qu'ils se furent appliqués à ces sons, ils en formerent d'articulés pour exprimer mieux leurs pensées. La raison corrigea la nature, & accomoda les mots à la signification des choses…La necessité où les hommes étoient de parler les uns aux autres, les obligea d'inventer des mots à proportion qu'on trouvoit de nouvelles choses... Ce fut la raison pourquoi il fallut inventer de nouveaux mots, lors qu'on bâtit cette fameuse Tour de Babylone: & on ne doit pas s'étonner s'il y arriva tant de confusion, d'autant qu'il se présentoit quantité de choses qui n'avoient pas encore leurs noms. Chacun les exprimoit à sa maniere; & comme la nature commence ordinairement par ce qui est de plus simple & de moins composé, on ne peut pas douter que la premiere Langue n'ait été tres-simple & sans aucune composition.>>

"Diodorus of Sicily explains the invention of languages as follows. Men, in making their first attempts to speak, initially pronounced sounds that signify nothing: then, after applying themselves to these sounds, they formed articulations to express their thoughts. Reason corrected nature & adapted the words to the significance of things…The need for men to speak to one other obliged them to invent words in proportion to their finding new things…This is the reason why they were required to invent new words while building the famous Tower of Babylon & one should not be surprised if there was so much confusion because they were presented with so many things that didn't yet have their names. Each one expressed himself in his own manner & because nature ordinarily starts with what is the most simple and the least complex, one cannot doubt that the first Language was not simple and without complexity."

– Richard Simon (b. 1638 – d. 1712), Histoire Critique du Vieux Testament (1678), trans. French, on the dangers of miscommunication associated with complex technical innovation

We live in a world of increasing technical complexity. Large-scale engineering systems now dominate the landscape of our society. Examples of such system include multi-modal transportation, military acquisitions, and health care delivery, touching upon just about every domain of modern human experience.

As technical complexity increases, organizational complexity must necessarily follow since the detailed operations of the system begin to exceed a single human's cognitive capacity (Conway 1968). More individuals will therefore be required to construct, maintain, and understand the systems upon which we rely for our way of life. The communication and aggregation of relevant knowledge among these individuals could conceivably benefit from an explicit design effort to ensure that the right knowledge is possessed by, and shared among, the appropriate people.

A traditional engineering organization in the United States typically responds to complexity via specialization. In other words, individuals are trained, recruited and paid to focus on a particular subsystem. Knowledge of, and experience with, the inner workings of system components is spread among expert specialists. Any large-scale engineered system must also receive the approval of several stakeholders of the system and its functionality, many of whom have different perceptions and hence, different requirements. Examples include design reviews that large-scale engineered systems must pass (consider, for example, the PDR and CDR cycles within the aerospace domain). These approval activities bring additional expertise to bear on improving the ultimate design. Highly experienced specialists develop expertise which is then communicated to mid-level managers, who are responsible for aggregating experts' recommendations and passing this information to upper-level management. For this procedure to work, problems faced by the decision-making organizations must be quickly diagnosed as relevant

to a particular specialty. The appropriate specialist must then possess the correct knowledge if s/he is to make a decision that is in the best interests of the organization.

Different experts, having been trained in different areas or components, will tend to pay attention to those elements of the system that they find consistent with their professional training – i.e., cognitively salient (Douglas 1986). The mechanisms by which this training is achieved include acculturation within specific professional specialties, and require learning that professional institution's language and jargon. By institution, we mean a set of social norms to which a particular community adheres. This leads to a situation wherein individual experts develop different views of the system. In such cases, the system becomes a boundary object (cf. Carlile and Schoonhoven 2002), knowledge about which must be jointly constructed by the experts in question.

In the committees that concern us in this thesis, information must be aggregated from multiple expert specialists. Evaluating committee decision processes requires a means of understanding the interaction between the social and technical specifics of the system in question. For example, (Jasanoff 1987) notes that disagreements between technical experts regarding whose expertise is most appropriate might lead to "boundary conflicts", wherein each expert attempts to define the problem in such a way as to make it consistent with his or her realm of expertise. The decision of what information is important and how it should be interpreted is the subject of exchange up until the time that each committee member casts a vote. That different experts hold different perspectives and values makes it more likely that additional aspects of a problem will come under consideration. Nevertheless, this does not guarantee consensus on the interpretation of data.

There is much evidence to suggest that decisions that are informed by a diversity of viewpoints are superior (e.g., Hong and Page 2004). This is because different experts will bring different domains of knowledge to bear on solving the problem at hand, potentially leading to a better-informed decision outcome. Modern technical organizations therefore require a capacity for lateral (i.e., non-hierarchical, or informally hierarchical) communication, especially if the organization is to respond quickly to uncertain future states (Galbraith 1993). Nevertheless, with specialization comes acculturation – we have noted that specialists, having been differentially trained, view the system differently. Thus, with acculturation may come difficulty in communication across specialty boundaries. This might be due to disagreement on goals (i.e., local vs. global optimization) or a simple inability to comprehend the jargon of specialists from other disciplines. This motivates three main questions driving our research endeavor:

1. How can we study, in a repeatable, consistent manner, the flow of communication among technical experts in committee decisions?

2. How do technical experts' decisions change as they learn and interact during the decision-making process?

3. How might we design committee procedures so as to enable desirable behavior on the part of technical expert committees?

The question of how to design decision-making processes that successfully leverage different perspectives is one that is extensible to a range of technology and policy activities across both public and private sectors. We differ from previous analyses in our use of an empirical quantitative methodology based upon analysis of meeting transcripts. Such a methodology can be extended to similar studies in other domains of interest to engineers, managers and social scientists.

This thesis presents an empirical method aimed at extracting communication patterns through a computational analysis of committee meeting transcripts. A computational approach is used for its consistency and reliability across meetings. Furthermore, an algorithmic approach enables any potential biases that might be present in the analysis to be minimal and transparent. Indeed, the process of developing such a methodology is an exercise in making such biases explicit and then, systematically attempting to eliminate those that are unjustified.

In particular, we use a modification of the Author-Topic Model (Rosen-Zvi et al. 2004), a Bayesian inference tool used in the field of machine learning to discover linguistic affinity between committee members. We find that the resulting output may be used to construct social networks representing patterns of communication among panel members. Analyses of these networks are then performed. Finally, a computational model is constructed that points the way forward for theory development.

Thesis Outline

Decision-making by groups of experts is an area that touches on a number of different disciplines within the social sciences. In Chapter 2, we review the literature on decision-making in small groups across economics, political science, social psychology, sociology and anthropology. This review motivates the need for a methodology that can analyze real-world decision-making by committees of experts.

Chapter 3 identifies the FDA Medical Device Advisory Panel committees as a relevant data source and reviews the structure of the panel decision-making process.

Chapter 4 introduces an empirical methodology which generates directed social networks from meeting transcripts based on Bayesian Topic Modeling.

Chapter 5 analyzes the networks generated from a set of 37 meetings of the Circulatory Systems Devices Panel and presents empirical findings.

In Chapter 6, we present a computational model that embodies a theory of panel voting procedures. Model results are presented and compared to empirical results, and directions for future modeling work are outlined.

Finally, Chapter 7 concludes by drawing implications from the empirical and modeling results for the design of expert committees in engineering systems.

C h a p t e r 2

LITERATURE REVIEW

"Siquidem pene totum humanum genus ad opus iniquitatis coierat: pars imperabant, pars architectabantur, pars muros moliebantur, pars amussibus regulabant, pars trullis linebant, pars scindere rupes, pars mari, pars terra vehere intendebant, partesque diverse diversis aliis operibus indulgebant; cum celitus tanta confusione percussi sunt ut, qui omnes una eademque loquela deserviebant ad opus, ab opere multis diversificati loquelis desinerent et nunquam ad idem commertium convenirent. Solis etenim in uno convenientibus actu eadem loquela remansit: puta cunctis architectoribus una, cunctis saxa volventibus una, cunctis ea parantibus una; et sic de singulis operantibus accidit. Quot quot autem exercitii varietates tendebant ad opus, tot tot ydiomatibus tunc genus humanum disiungitur; et quanto excellentius exercebant, tanto rudius nunc barbariusque locuntur."

"Virtually all of the human race had united in this iniquitous enterprise. Some gave orders; some did the planning; some raised the walls; some straightened them with rule and line; some smoothed mortar with trowels, some concentrated on cutting stone and others on transporting it by land and sea. Thus diverse groups applied themselves in various ways, when they were struck by Heaven with so great a confusion that though all had been using the same language in their work, made strangers to one another by the diversity of tongues, and never again succeeded in working together. Only each group that had been working together on one particular task kept one and the same language: for example, one for all the architects, one for all the stone movers; for all the stone-cutters, and so on with every trade. And now as many languages separated the human race as there were different kinds of work; and the more excellent the type of work, the more crudely and barbarically did they speak now."

– Dante Alighieri (b. 1265 – d. 1321), De Vulgaris Eloquentia (1302-1305), Book I, Section 7, 6-7. trans. Latin, M. Shapiro (1990), on the origin of language and technical jargon

Chapter 1 noted the necessity for information aggregation in order to make decisions within engineering systems. The primary means by which this occurs is via the expert committee – a body charged with the review and approval of complex projects. The committee is a common means by which experts pool their knowledge in an attempt to reach a consensus decision about a complex system or process. A successful committee will be able to integrate the disparate knowledge and viewpoints of its members so as to make a decision that is as well-informed as possible (moderate success will involve coming to a better decision than a randomly informed decision maker). An unsuccessful committee can fail in a number of ways – for example with decisions that are less than optimal but still quite good, or with very poor decisions. Lack of success can hold for many reasons. These include, but are not limited to, the absence of relevant technical expertise; the inability of committee members to communicate (e.g., across disciplinary boundaries); and personality conflicts (see Boffey 1976, for an example of these challenges in the early FDA drug and medical device approval committees).

Pooling and communication of information across specialty boundaries is not straightforward. Committees are social entities and are therefore affected by a number of mechanisms recorded in the social sciences. Our challenge is to determine which of these are likely to be encountered by committees of technical experts and to evaluate how they might impact upon decision outcomes. To do this, we review a number of bodies of academic literature that are concerned with committee decision-making, motivating the methodological approach outlined in Chapter 3.

Rational Choice

A natural place to begin a review of literature on committee decision-making is in the rational choice tradition where we find the most precise theoretical constructs

due to well-developed mathematical machinery. This literature typically follows the utilitarian philosophy of Jeremy Bentham, arguing that individuals seek to maximize their individual utilities in any decision-situation (Bentham 1988/1780). This individual behavior then aggregates to often-unexpected group behaviors. The advent of utility theory by von Neumann and Morgenstern (2007/1947) provided one mathematical framework for this approach which was ultimately developed into the body of literature now known as Game Theory. Economists and political scientists have applied this theory to committee decision problems, most notably those of bargaining, (Rubinstein 1982), organizational structure (Sah and Stiglitz 1988), coalition formation (Aumann and Dreze 1974; Baron and Ferejohn 1989), and recently, ritualized common knowledge (Chwe 2003). These models are highly appropriate for their political science context (e.g., in legislative bodies) where individuals frequently behave in a manner aimed to maximize the interest of their own constituents, even when cooperation and exchange of expertise is required (cf. Broniatowski and Weigel 2006). Therefore, a major focus of this work is strategic and payoff-focused, and therefore does not incorporate to the technical specifics of the question that the committee is considering. Furthermore, decision-makers are treated as homogeneous, whereas decision rules are instead made to vary. More recent work has begun to focus explicitly on committees of experts (Visser and Swank 2007). Although theoretically compelling, this work has not yet incorporated the notion that experts may be qualitatively different from one another (e.g., with access to different sources of information or different value structures). This work recognizes different levels of ability among experts, but not different kinds of knowledge or expertise. Such an approach could therefore be augmented by an incorporation of heterogeneity among the knowledge and beliefs of decision-makers.

Social Choice

Social choice is a second strand of modeling literature within the traditions of economics and political science. Unlike rational choice, social choice does not require algebraic utility functions, instead relying on simple ordering among preference alternatives (Gaertner 2009). Assuming the existence of at least three alternatives, Arrow (1963) demonstrated the logical impossibility of generating a stable preference ranking under a set of conditions that have been generally accepted as reasonable namely:

1. Non-dictatorship: The preference ordering for the group must not depend solely on the preference ordering of only one individual.

2. Unrestricted domain: All preference alternatives of all individuals should be included in the ranking, and no comparison among pairs of alternatives is invalid *a priori*.

3. Independence of Irrelevant Alternatives: Changes in an individual's preferences between decision alternatives that are not included in the group preference ordering will not change the group preference ordering.

4. Monotonicity: An increase (decrease) by an individual in his/her preferences such that one alternative is ranked higher (lower) than it had been previously can not result in a lower (higher) ranking for that alternative in the group preference ordering.

5. Non-imposition: Every group preference order could be attained by some set of individual preference orders.

Arrow's theorem would seem to suggest that a rational (i.e., non-cyclical) aggregated preference ranking is impossible. Nevertheless, if one were to impose a domain restriction on the preferences of individuals (i.e., relaxation of

assumption number 2) one could generate stable preference orderings (Black 1948). One possible source of a domain restriction could be "empirical reality" (Frey et al. 2009), which would prevent certain inconsistent interpretations of a given decision-situation, and hence, certain preference orderings. In principle, as more information becomes available to committee members, they should converge on the right choice. Indeed, consensus in the presence of large amounts of information might be one definition of expertise (cf. Romney 1986). Work performed by Whitman Richards and his colleagues has shown that creation of a "shared knowledge structure", i.e., a socially shared set of relations among preference alternatives, greatly increases the likelihood of a stable preference ordering (Richards et al. 2002). This suggests that the creation of a shared interpretation of the data representing a given system under analysis is critical for the committee to reach some kind of agreement, even in the absence of common preferences. Even without definitive data, committee members from the same specialty or discipline might share such an interpretation *a priori* due to commonly held assumption, beliefs and training (Douglas 1986). Furthermore, as a committee becomes more diverse, we would expect more possible interpretations to become available. In the absence of a shared interpretation the committee might be unable to agree – this should be particularly true when there are multiple possible interpretations and the data is sufficiently ambiguous to rule out few of them (March 1994). Furthermore, very complex systems may be most likely to lend themselves to multiple viable interpretations. In the absence of clear communication (e.g., learning across disciplinary boundaries), we should expect a tradeoff between a panel's diversity and its capacity to reach consensus. If communication is flawless, i.e., learning always occurs, agent-based modeling work by Scott Page has shown that, a diverse group that is individually less expert will outperform a homogeneous group of individuals who are each more expert (Page 2008). This is because each expert in the diverse group can bring a different set of perspectives and heuristics to bear upon solving a common problem. Once

that expert has reached his/her local optimum, another expert takes over. Underlying Page's model is an assumption that there exists a global optimum for all decision-makers. This assumption provides a domain restriction in the sense of Arrow's theorem that is consistent with the notion of an empirical reality. The assumption of perfect communication among committee members further supports this shared outcome; nevertheless, it is not always realistic. Page's work suggests that, if useful communication can be established in committees, we might expect better outcomes. We therefore turn to the empirical social psychology literature for insight into communication patterns in small groups.

Social Psychology

One of the foundational researchers in social psychology, Leon Festinger, outlined a theory of social comparison processes based upon the notion that individuals within a small-group setting constantly compare their performance with other group members (Festinger 1954). These comparisons are based on social and physical comparisons that depend on meeting context, e.g., demonstrated expertise in a committee meeting. The implication is that similar individuals may be driven to behave in a way that emphasizes their value to the group. As a particular trait or opinion becomes elevated in importance, there is a "pressure toward uniformity", when group members, in trying to demonstrate their value, become less willing to deviate from the opinions expressed by their peers. This is one explanation for the phenomenon commonly known as "groupthink". Festinger's theory has implications for expert committees because it suggests that individual members might not be willing to reveal important information if it would lead them to draw a conclusion differing from that espoused by the majority. Festinger further noted that social comparisons decreased in importance as similarity decreased between group members. This indicates a second benefit of diversity, suggesting that there are ways in which diversity promotes, rather than inhibits, communication.

Festinger's theory does not consider actual expertise and provides no information regarding whether it is equally plausible that committee members might converge on a correct interpretation of the data as on an incorrect one. Thus, we might ask the circumstances under which a panel that does reach consensus is likely to achieve the correct outcome. Experiments run by Bottger (1984) suggest that a distinction can be made between actual expertise and perceived influence. Using a simulated NASA mission, Bottger found that group members were most influenced by the correct statements of panel members – experts were not ignored. On the other hand, when asked to rate which group member was the most influential, other group members frequently identified those who spoke most often. Bottger concluded that groups often do not attribute influence correctly to their members. Furthermore, groups make the best decisions when actual expertise and perceived influence (i.e., air-time) covary. In the absence of this covariance, actual experts do not have the opportunity to contribute their useful knowledge. In other words, experts must be given the opportunity, and inclination, to express their views. Bottger's findings supplement Festinger's theory by providing information about task performance, while emphasizing the importance of focusing group efforts in the right direction.

A much referenced paper by Stasser and Titus (1985) showed that in situations in which individual students possessed different information regarding a group decision, conversation and recall were both dominated by shared information – i.e., group members tended to discuss information that everyone else already knew. The implications of this result for group decision-making suggested that specialized information, e.g., due to expertise, was unlikely to be shared, thus resulting in an outcome biased towards shared information. Not only did discussion fail to promote information exchange, it succeeded in promoting a biased viewpoint. Stasser and Titus (1987) confirmed this result in a second experiment which showed that unless group members shared very little

information to begin with (i.e., had very diverse information sets), shared information dominated the discussion and perpetuated the associated bias, even though individual group members had access to more information that contradicted the group's decision. This suggests that group diversity might lead to better communication. Stasser (1988) explained this result with a computational model, called "DISCUSS" which posited that individual group members sampled information items uniformly. Shared information was much more likely to be discussed simply because each group member had access to it. Using the DISCUSS model, Stasser showed that no explicit bias was necessary in order to explain his earlier results – i.e., group members did not have to behave strategically to create groupthink. In a second paper Stasser (1992) used DISCUSS to show that increasing the salience (i.e., probability of discussion) for an unshared item did not significantly change the probability of that item's being mentioned except for very small groups (four members or fewer). The results presented by Stasser and Titus would seem to suggest that learning among group members is unlikely, even in the absence of the social comparison processes identified by Festinger.

One major limitation of the early work of Stasser and Titus is its empirical reliance on samples of undergraduate students. In particular, this literature suffers from a problem of external validity. Although decision-makers are not assumed to be homogeneous as in the rational choice literature, they only differ in the information in which they possess. The experimentally-controlled knowledge shared among groups of undergraduates does not constitute domain expertise of the sort that we would expect among committees of technical experts. This suggests that the generalization of these findings to real-world scenarios might not hold. This motivates the need for an analysis of actual group decisions by technical experts, rather than experiments run on groups of students. Furthermore, no group member is aware of the knowledge possessed by other

members. A seminal paper by Wegner (1987) introduced the concept of "transactive memory" – the notion that a group requires meta-knowledge in order to perform efficiently. Group members must know who else in the group holds what knowledge. When transactive memory is present, group members are able to assign specialists appropriately and learn from one another. It stands to reason that this capacity would be of particular value on a committee of technical experts. Indeed, even when groups of students were tested, Stasser et al. (1995) found that the public assignment of expert roles led to the sharing of otherwise unshared information. Stasser et al. attributed the success of their scheme to a "cognitive division of labor" of the sort described by Wegner. This paper overturned the probabilistic conception of information sharing found in (Stasser 1992), but did not tender a new computational model to explain it. It is interesting that the findings of Stasser et al. (1995) are consistent with Festinger's social comparison theory – in particular, individuals who have publicly recognized expert roles may be perceived, and may perceive themselves, as different from other panel members. This would decrease the strength of their "pressure toward uniformity". It further suggests that that, on expert committees, each expert should be assigned a public role consistent with the knowledge that that expert is expected to convey. Although decision-making groups might be *ad hoc*, the roles of their members should not be. It seems possible that the association of experts with known specialties may provide this function in real groups. For example, public knowledge of the biographies of other panel members would promote meta-knowledge of the sort described by Wegner.

Acknowledging the external validity problems inherent in generalizing from laboratory conditions to real-world expert committees, the findings of Stasser et al. (1995) are encouraging if the expertise necessary to solve a problem is known. In cases of deeper uncertainty where there is no objective standard of expertise, a group might still be able to reach a consensus. Kameda et al. (1997) show that the

key to this is "cognitive centrality," a concept related to breadth of expertise. A group member is cognitively central if they share at least one information item with many other group members. This implies that their knowledge is socially validated. Such group members come to be viewed as credible sources of expertise by others, thereby creating an effect very similar to pre-existing meta-knowledge. Kameda et al. showed that because cognitively central members possess more shared information, they are more often able to change the preferences of their peers. They are also more resistant to preference change under influence from others. The relation between shared information and perceived expertise is further confirmed by Winquist and Larson (1998) who propose a dual-process model in which individuals discuss shared information so as to build their perceived expertise, whereas changes in actual preference only occur as a result of discussion of unshared information. This finding mirrors that of (Bottger 1984) with perceived expertise corresponding to the discussion of shared knowledge and actual influence corresponding to the discussion of unshared knowledge. Thomas-Hunt et al. (2003) suggest that this dynamic might also be linked to social validation with their experiments showing that socially connected members tend to focus more on shared information, whereas socially isolated members tend to focus more on unshared information. Furthermore, socially-connected group members evaluated the contributions of others positively when they followed this scheme and negatively otherwise. Although actual influence and perceived expertise are separate, it seems that concerns about social validation in small groups may cause perceived expertise to drive individual behavior. These results also tend to confirm Festinger's social-comparison theory in that individuals are more likely to mention a shared topic if other, comparable, group members have done so. The implication for committees of experts is that, when possible, clear roles should be assigned to committee members so as to avoid these effects. In the absence of clear knowledge regarding which sources of expertise are appropriate for decision-making (e.g., under conditions of high

ambiguity), group members with a breadth of expertise, able to validate the knowledge of others via the expression of shared information, are likely to serve a key role. Shared information, therefore, serves a similar role as a "shared knowledge structure" (Richards et al. 2002). This suggests that breadth of expertise might be particularly valuable under conditions of ambiguity because it might enable more knowledge sharing to a wider population. If group members are very specialized, they will be unable to communicate with one another. On the other hand, if a group member can communicate across specialties, deeper information that was otherwise unshared might be shared after the broad expert first mentions it. Furthermore, the broad member's expertise is recognized by other group members, giving him/her significant influence. However, if there are too many group members that are too broad, there is likely to be a large amount of shared information. This may create an incentive for these members to focus on the information that they share in common at the expense of the valuable unshared information that they might otherwise elicit from other group members who possess a depth, rather than a breadth, of expertise.

Sociology of Small Groups

We note that certain members are more likely to speak than are other members. A cognitively-central member of a group will likely be setting the agenda, although this might occur through a focus on knowledge that is already shared. Ideally, an expert would identify and discuss relevant topics and issues, and members who are less expert would follow. If the members who utilize the most air-time are not experts, it not impossible that the necessary expertise might not be revealed. Within the literature on the sociology of small groups, propensity to speak is commonly referred to as "status". Given Bottger's finding that perceived and actual expertise must covary in order to generate a well-informed meeting outcome, a deeper understanding of the determinants of status would be

enlightening. We therefore turn to the literature on the sociology of small groups in order to better understand status effects in expert committees.

Expectation States

Foundational work by Báles et al. (1954) found that unstructured small groups (three to ten people) consistently generated hierarchies among group members. This finding was robust across a range of domains and groups. This manifested as a strict ranking where the person at the top of the hierarchy spoke most often and was most often addressed by others. The next person was ranked second in terms of his/her total number of utterances generated and received, etc. This finding is quite robust within the sociology literature, and perhaps reflects the distribution of perceived influence as in Bottger (1984). If such is the case, then actual expertise should be made to covary with these characteristics through the institution of procedures that embody this perceived expertise. This raises the question of how to determine *a priori* the hierarchy that seems to emerge organically in a small group. Hare and Bales (1963) found that status is often correlated with physical location in a group, such that members who are central display a higher propensity to speak, and that members who are physically distant will be less likely to interact. Furthermore, personality testing showed that when seats were not assigned, more dominant personalities tended to choose more central seating locations. Such results suggest that attention paid pre-meeting to procedural variables, such as seating location, could strongly impact on the information that is shared and, consequently, on the ability of the panel to successfully pool information.

Other research within sociology has found that the status hierarchies identified above are associated with the personal attributes of speakers, including their age, race, and gender. In a foundational paper for what came to be known as the "expectation states" paradigm in small group research, Berger et al. (1972)

provided an extensive overview of the literature on status characteristics in small groups, and then proposed a theory to account for it. Berger et al. propose status as an explanatory variable which determines "…evaluations of and performance-expectations for group members and hence the distribution of participation, influence, and prestige." (Berger et al. 1972). Furthermore, Berger et al follow Festinger in noting that "a status characteristic becomes relevant in all situations except when it is culturally known to be irrelevant" (Berger et al. 1972). Status is therefore proposed as an abstract hidden variable which explains the hierarchies identified by Báles et al. Skvoretz (1981) extends expectations states theory with a mathematical formulation of this concept based upon data from a psychiatric hospital. He identifies two dimensions of status – namely, position in the hospital hierarchy and clinical competence. It is interesting that, in this context, the two dimensions of status are unrelated to age, race and gender, as specified by Berger et al., and are instead related to expertise. Recognizing that external social relations often contribute to the formation of status hierarchies within groups, Fararo and Skvoretz (1986) introduced "E-state structuralism" – a mathematical synthesis of the expectation states literature outlined above and the structuralism of the social network literature in order to explain the change over time of dominance relations in small groups, including groups of animals. Smith-Lovin et al. (1986) tested Skvoretz's (1981) mathematical formulation and found that it did not explain the participation rates of six-person task oriented groups of undergraduates that were explicitly designed to vary along the dimension of gender. This is because they found a large degree of variation within gender groups. They proposed a two-dimensional refined model that first segmented each group by gender and then, within each gender, explained status characteristics using Skvoretz's approach. Skvoretz (1988) further tested this finding by systematically varying the gender composition of six-person groups, finding that none of his models sufficiently explained his data. These results would seem to indicate the importance of gender as a status characteristic in small

groups. Fişek et al. (1991) reviewed the data collected by Skvoretz (1988) and Smith-Lovin et al. (1986), noting an "undeniable gender effect" and introduced a mathematical model based upon expectation-states theory and the presence of external status characteristics. Nevertheless, it is important to note that Smith-Lovin et al. (1986) did not examine groups of experts as did Skvoretz (1981). Indeed, there seemed to be little else that could differentiate these undergraduates from one another since all other potential status characteristics were controlled for. These results suggest that, only in the absence of an existing hierarchy, such as that defined by the social structure of a hospital, or by mutually recognized expertise, might gender be an adequate explanatory variable. This interpretation is consistent with Festinger's notion that group members will differentiate themselves on the basis of characteristics that are relevant to the task at hand (1954).

Conversation Analysis

Parallel to the expectation-states literature is conversation analysis – a tradition that traces its roots to the ethnomethodology of Garfinkel (e.g., 1984), the observations of Goffman (e.g., 1981) and the work of Sacks, Schegloff and Jefferson (e.g., 1974) and is focused on generating a qualitative understanding of how the unspoken rules of conversation drive the content communicated. Maynard (1980) identifies topics as a key feature of conversations, arguing that changes in topic are non-random occurrences that can be related to the structure of the group that is discussing them. Okamoto and Smith-Lovin (2001) extend the insights of conversation analysis into the expectation states literature, with a focus on how status characteristics, of the sort identified above, impact on an individual's capacity to change the topic of conversation. Gibson (2003) notes that because only one person can generally speak at a time, external status characteristics manifest in conversation as participation-shifts and often as topic shifts. Gibson (2005) verifies this statistically, by linking network structure to

participation shifts. This shows a mechanism by which external status characteristics impact upon what is discussed, and is perhaps the first quantitative result in the conversation analysis tradition. These results suggest that an analysis of the speech of committee member participants might provide some deep insight into the dynamics of group decision-making, even among committees of experts.

Appreciative Approaches

The above literature still assumes that group members can perfectly understand information that has been communicated, regardless of its source. If such is the case, then a correct structuring of a given committee should, as Page predicts, lead to selection of the best solution to the problem being studied given the information available. Nevertheless, literature relating language and culture suggests otherwise. This literature notes that language, beyond being simply a means of exchanging information, is also an expression of identity. This idea began as a philosophical concept, perhaps most strongly connected with the German Idealists (von Herder 2002/1767, von Humboldt 1997/1820), who argued that national language both reflects culture and shapes patterns of thought. According to this tradition, different national languages represent different world-views that are incommensurable. The implication is that it is impossible to truly translate from one language to another – some element of the original concept must be lost, a question of great interest to literary criticism (e.g., Benjamin 1969; Eco and McEwen 2001). This idea was formalized within linguistics as the Sapir-Whorf hypothesis – the notion that linguistic structure and usage places limits on the cognition of its users (Sapir and Mandelbaum 1947; Whorf et al. 1998). Although the Sapir-Whorf hypothesis had fallen out of favor in linguistics, it has begun to be rehabilitated by the work of Lera Boroditsky – a cognitive scientist who has tested the effects of Chinese and English language on cognition (Boroditsky 2002).

Work within the anthropology and Science, Technology and Society (STS) literatures extends this notion to the realms of professional and institutional cultures. In particular, the penetrating analyses of Mary Douglas note that group membership may affect perception of data (Douglas 1986). Membership is conferred upon those individuals who group features of the world into categories that are consistent with group norms. This is reflective of a wider principle in anthropology that different professional or institutional cultures will selectively direct individuals' attention to the elements that are salient within their group structures. Among technical experts, this is reflected in the fact that each specialty possesses its own unique language and jargon, which carries with it an implicit scheme for categorizing perceived phenomena (Brown 1986). On the other hand, an outsider to the group, who is unfamiliar with the jargon used, may be unable to understand the discourse. This is because the specific jargon refers to commonly held sensory and social experiences that a member of another institution is unlikely to have directly encountered. This is particularly true in medical and academic disciplines, where conceptual precision is required to communicate within the specialty. One could argue that just as Whorf's Eskimo has many different words for snow, and therefore a deeper capacity to categorize these types, a technical specialist has jargon that is specific to their specialty and, most importantly, to their experience. Communicating this experience is a classic dilemma in the knowledge management literature, largely because many of the important elements are tacit (Polanyi 1958; Nonaka et al. 2000). Nelson notes the importance of written and oral language as a means of encapsulating and transferring tacit knowledge (Nelson 2005). On the other hand, an outsider to the institution may be unable to understand the discourse because they lack the underlying shared experience. The STS literature extends this notion by noting that language is used as a cognitive mechanism to delineate professional boundaries. This directs the attention of experts within a specialty toward a given interpretation of a problem that is consistent with that expert's training, while

simultaneously directing that attention away from other possible interpretations (Cohn 1987; Mulkay et al. 1987; Rapp 2000; Winner 1986). The same groups that drive selective perception and word choice also confer a sense of identity. March notes the dialectic between decision-making as rational choice based on a consequentialist pursuit of preferences and identity-based rule-following, ultimately noting that each viewpoint supplements the other, particularly under conditions of low ambiguity when roles are clear (March 1994). We may therefore expect preferences to be correlated with group membership and, by extension, its associated jargon. Furthermore, this literature suggests that even when each speaker is given the appropriate amount of time to discuss their viewpoints, a listener might have trouble understanding or assimilating all of the implications of the information being shared. Communication within a group should be relatively easy, whereas communication across groups may be more difficult. This strongly motivates the presence of at least one interdisciplinary panel member who can act as a translator so as to be able to assimilate information from members of other specialties, and as a teacher, so as to be able to communicate with members of the same specialty.

Strategic Behavior within Groups

In addition to the difficulties in comprehension that might exist across group boundaries, literature in political science and STS suggests that group loyalty could possibly lead individuals to focus on group goals over those of the committee as a whole. Casting "organization [as] the mobilization of bias", (Elder & Cobb 1983) recognizes institution-specific symbolism in language, noting that the choice of terminology in defining a problem may be seen as a means of mobilizing support. Furthermore, the linguistic definition of a problem dictates, to some extent, its solution. Choosing to use specialized technical words serves to narrow the range of subjective meaning of otherwise ambiguous terminology (such as "safety" or "efficacy" in FDA's context) thereby implicitly redefining the

problem according to a given speaker's particular interest. This can be viewed as an example of "agenda-setting" behavior, which is common in political discourse. Here, the major concerns are related to framing and issue-redefinition (Cobb and Elder 1983). In the worst case, if a given interpretation of data is favored by one group over another, that group may try to promote a particular interpretation rather than attempting to learn from one another. This worst-case scenario would be an example of contested boundaries among groups of experts in science policy (Jasanoff 1987). However, this sort of behavior need not be strategic among committees of experts and could instead be a way of expressing the knowledge inherent in a particular specialty's training scheme. Experts may try to learn from one another even though they state their positions. Learning could become a particularly difficult problem under conditions of ambiguity (March 1994; Lawson 2008). Furthermore, the more a given pair of interpretations are equally plausible, the more likely that the committee will be unable to rule one of them out, potentially leading to a preferential selection according to group membership. In such a case, learning may not occur – instead the majority group within the committee may decide policy. Douglas and Wildavsky (1982) further note that different individuals are likely to have different values. Under conditions in which group norms are clear, these values may not be evident. March notes that under conditions of ambiguity, values (e.g., sacred values, cf. Tetlock 2003) become more important in driving behavior (March 1994). Even if individuals share the same perspective, agreement on an appropriate course of action may be unlikely if questions of values or identity (e.g., ethical or other moral dilemmas) are involved. This last point is particularly important for complex engineering systems, in which lives and livelihood frequently depend on the correct operation of the system.

C h a p t e r 3

CASE STUDY: FDA ADVISORY PANELS

דיני ממונות הטהרות והטמאות, מתחילין מן הגדול; ודיני נפשות, מתחילין מן הצד.

"…in legal matters involving money, and ritual purity and impurity, we begin with [the opinion of] the greatest [of the judges]; whereas in legal matters involving capital charges, we begin with [the opinion of] those on the side [benches]…"

– Mishna Sanhedrin 32a, trans. Hebrew.

"According to the Oral Tradition, we learned that with regard to cases involving capital punishment, we do not ask the judge of the highest stature to render judgment first, lest the remainder rely on his opinion and not see themselves as worthy to argue against him. Instead, every judge must state what appears to him, according to his own opinion."

– Rabbi Moshe ben Maimon (Moses Maimonides), b. ca. 1137 – d. 1204, Hilchot Sanhedrin V'HaOnshin Hamesurim Lahem, (The Laws of the Courts and the Penalties placed under their Jurisdiction), Chapter 10, Par. 6, trans. Hebrew, Rabbi S. Yaffe, on the impact of procedure in group decision-making

The U.S. Food and Drug Administration (FDA) advisory panel meetings are a committee evaluation of a complex engineered system involving different specialties. Furthermore, transcripts of these meetings are generated by a court-recorder and are available to the public as stipulated by the American Federal Advisory Committee Act. The transcripts of these panel meetings therefore provide a rich data source from which we may study technical decision making by committees of experts (Sherman 2004). Decisions made by technical expert committees in the FDA are analogous to those that must be made by committees of technical experts within other complex engineered systems. As explained above, different experts may possess varying interpretations of data, potentially leading to alternate, but equally legitimate, readings of uncertain evidence. Reaching a design decision requires that information from these different specialties be aggregated in some way. Ideally, the ultimate decision would be well-informed by all perspectives in the room.

Multi-Stakeholder Environment

As in a decision involving multiple stakeholders in a complex engineered system, the FDA decision-making process is embedded in a policy environment. The task of approving medical devices for the US market falls to the Food and Drug Administration's Center for Devices and Radiological Health (CDRH). CDRH classifies each device into one of the classes, grouped by risk (see Figure 1).

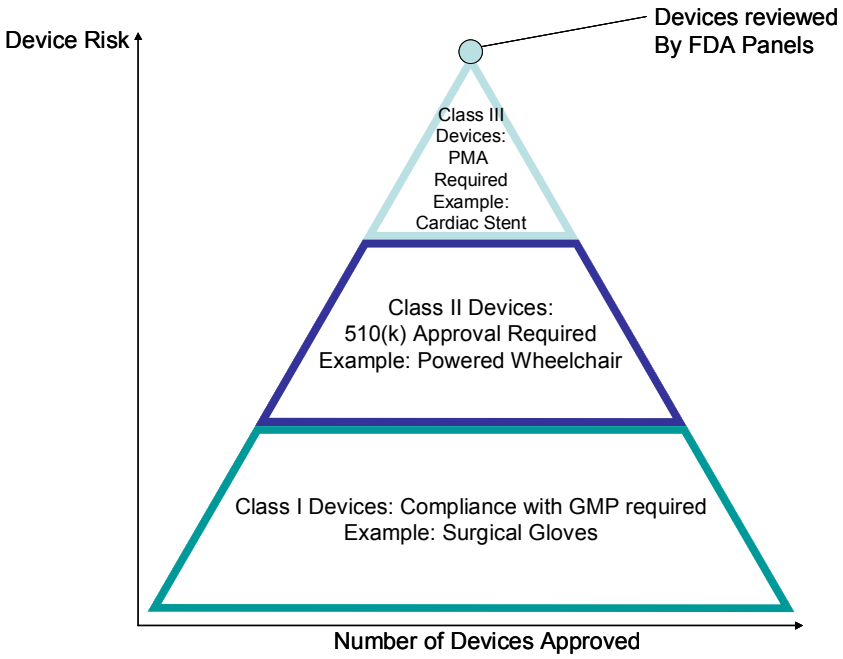


Figure 1: Medical devices are classified into three categories based upon risk to the patient. PMA = Pre-Market Approval; GMP = Good Manufacturing Practices.

Figure 2, sourced from (Maisel 2004), provides an overview of the process by which a device is reviewed for approval by CDRH.

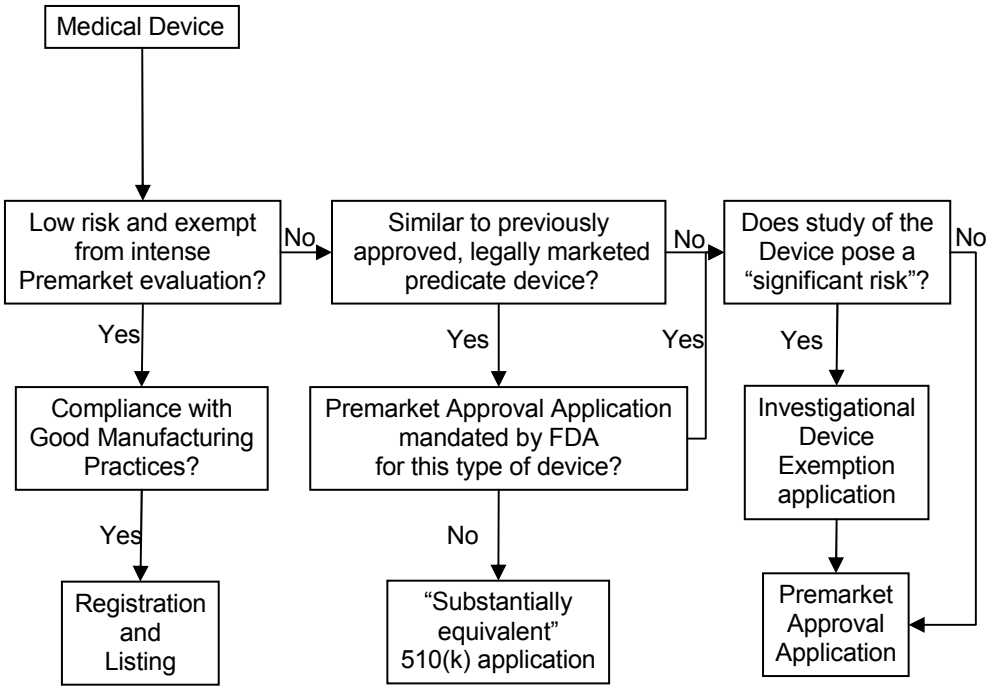


Figure 2: Medical devices are classified into three categories based upon risk to the patient. Diagram sourced from (Maisel 2004).

The grant of a 510(k) or Pre-Market Approval (PMA) by the FDA allows a device to be marketed in the United States. These approvals often act as *de facto* monopolies for the device involved because any competitor must demonstrate additional safety or efficacy of the new device as compared to the initial baseline in order to receive approval. Advisory panels review devices "as needed" (Parisian 2001). Devices brought to committees for review are generally those which the FDA does not have the "in-house expertise" to evaluate. As such, the devices under evaluation by the committees are likely to be the most radical innovations facing medical practice, and those facing the most uncertainty. Furthermore, advisory panel members are "by definition, the world's experts who are engaged in cutting-edge bench science, clinical research and independent

consulting work" (Sherman 2004). Advisory panels therefore serve to bring needed expert knowledge and political credibility with industry and consumer advocate groups to the FDA device approval process. In practice, a very small proportion of all devices submitted to FDA for approval are reviewed by the panel. Audience members will include representatives of the media, consumer advocate groups, the financial community, and competitor companies, all of whom are looking for information regarding how the medical device might perform on the market (Pines 2002). Therefore, panel recommendations and the judgments and statements of individual members carry significant weight both inside and outside the FDA.

Panel Procedures

A typical FDA advisory panel meeting follows a number of sequential procedural steps, as follows:

1. Introduction and Conflict of Interest Statement

In this part of the meeting, each panel member is introduced and the Executive Secretary reads a statement regarding the degree and source (sponsor or competitor) of a given panel member's potential financial conflict of interest. Panel members are seated at a U-shaped or V-shaped table with the committee chair located at the apex. A representation of this arrangement is shown in Figure 3 (FDA 1994).

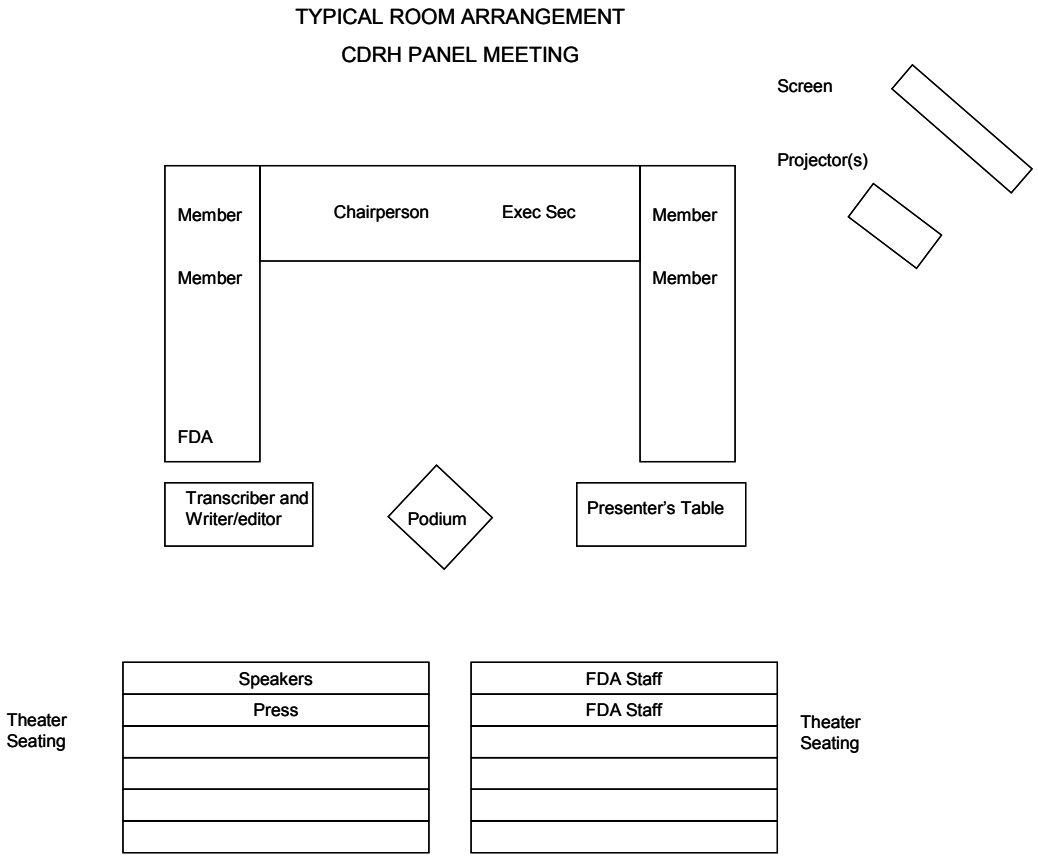


Figure 3: Standard Layout for FDA CDRH Advisory Panel Meeting.

2. First Open Public Hearing

In this stage, any member of the public can make a presentation to the panel.

3. Sponsor Presentation

In this stage, the device sponsor (usually a medical device company) presents their review of clinical trial results to the panel. Individual panel members may ask questions at the discretion of the committee chair.

4. FDA Presentation

In this stage, the FDA review team presents their review of clinical trial results to the panel. Individual panel members may ask questions at the discretion of the committee chair.

5. Panel Questions

In this stage, the panel members ask questions to the sponsor and FDA representatives. The phase will often begin with presentations by one or two panel lead reviewers, followed by questions asked by panel members at the discretion of the chair. In practice, the committee chair typically chooses one person to begin asking questions (e.g., sitting adjacent to a lead reviewer, or sitting at one end of the table). Other panel members proceed to ask questions in order around the table.

6. Open Discussion

After each panel member has had an opportunity to ask questions of the FDA and sponsor, there is an open discussion session in which each panel member may ask additional questions and discuss the device application. This discussion is often guided by questions asked to the panel by the FDA Executive Secretary regarding recommendations for approval.

7. Open Public Hearing

A second open public hearing is held to allow members from the public to speak.

8. Panel vote

Panel members move for approval, approval with conditions, or non-approval of a device. Although panel members might bargain over which conditions of

approval to include, this often occurs implicitly during panel discussion. Panel members then vote in an order determined by the executive secretary (i.e., sequentially or simultaneously). Finally, panel members express their reasons for their votes. The committee chair then adjourns the meeting. The meeting typically has breaks for lunch between stages three and four or between stages four and five. In addition, the committee chair can call for a coffee or restroom break at his/her discretion.

Within the health-care domain, there has been a movement towards "evidence-based medicine" (Sackett et al. 1996). Although aimed at integrating clinical expertise with experimental findings, this movement has often been interpreted by practitioners as privileging population-level experimental results over the expertise of individual practitioners. It is in response to this narrow definition that Gelijns et al. (2005) note that decisions cannot be strictly "evidence-based" for the following reasons:

1. A given data-set may be interpreted differently by different experts, especially in the presence of high uncertainty. Unless these experts can learn from one another, good decision-making might be impaired.

2. Patterns of technological change are difficult to predict, particularly when innovations are ultimately used for different purposes than originally intended.

3. Even in the case of clear evidence, decision-makers may disagree on its implications due to differing value systems.

These are all reasons why expertise must be integrated with evidence; not replaced by it. Although referring to health care, these caveats apply equally to any engineering system. Unless experts can learn from one another, good

decision-making will be impaired. This suggests that a device's determination as safe or efficacious depends strongly on factors that are not within the purview of "evidence-based" decision-making, narrowly defined.. Douglas (1986) argues that these are largely shaped by the perceptions, and hence, the knowledge and expertise, of risk assessors. Groups that might impact decision-making include membership in a particular profession, specialty, or bureaucratic organization. When combined, the diversity of viewpoints arising from these different groups may lead to a better decision outcome than that reached by one limited interpretation of the evidence.

Collaborative Technical Decision-Making in the FDA

As in any complex engineered system, technical experts in the FDA may not have an explicitly political aim. Nevertheless, their decisions may be perceived as biased by those who believe they would have made a different decision in their place. Although FDA advisory committees are aimed at producing "evidence-based" recommendations, differential interpretation of the evidence allows room for debate, and concomitant accusations of bias. Panel members' professional experiences might allow for intuition that can seem to go against the indications shown by the data. (Friedman 1978) expressed a concern that this constitutes a form of "specialty bias," especially when multiple medical specialties are involved. On the other hand, this view presupposes that a reading of the data that is entirely uninformed by past experience is best, which obviates the role of expertise in advisory panel decision making. A distinction must be drawn between decision-making that is based on evidence and decision-making that is driven by one "orthodox" reading of the evidence. Others argue that financial conflicts of interest should be mitigated in advisory panels. On the other hand, a prominent study recently found only a minor correlation between conflict of interest and voting patterns with no actual effect on device approval (Lurie et al 2006).

Data availability

One of the primary advantages to using the FDA Advisory Panels as a case study is the availability of data. There are 20 different panels whose transcripts are recorded over a period of ten years. This leads to the possibility of examining hundreds of committee meetings – a sufficiently large number that generalizable findings may be inferred. If the study were to expand to include the drug-approval committees within the FDA, the number of cases upon which we could draw would number in the thousands[1]. Furthermore, all panel transcripts are transcribed by a court-recorder, ensuring a standard of quality that is admissible in a court of law.

The empirical analysis mentioned above requires data in the form of committee meeting transcripts. These are often not recorded in textual form, or are proprietary to the organization that commissioned the committee. We therefore turn to transcripts of expert committee meetings that are a matter of public record. The ideal data source must have the following attributes:

1. Analysis or evaluation of a technological artifact
2. Participation of multiple experts from different fields or areas of specialization
3. A set of expressed preferences per meeting(such as a voting record)
4. Multiple meetings, so as to enable statistical significance

These requirements are met by the Food and Drug Administration's medical device advisory panels.

[1] Transcripts of FDA committee meetings are open to the public and located at: http://www.fda.gov/AdvisoryCommittees/default.htm

Chapter 4

METHODOLOGICAL APPROACH

אברא כדברא

Transliteration: "Abra Cadabra."

"I will create as I speak", trans. Aramaic.

This thesis is aimed at developing a deeper understanding of how communication on committees of technical experts impacts upon multi-actor decision-making through an analysis of the language used by each speaker in the discussion. The most direct way of deepening our understanding is to attempt to cluster speakers by the co-occurrence patterns of words in their discourses. In particular, this thesis presents an empirical quantitative methodology based upon a computational linguistic analysis of meeting transcripts.

A major challenge to the use of linguistic data for the analysis of social behavior on expert committees stems from the strong assumption that such dynamics are entirely reflected in language, and that differences in language necessarily indicate differences in perception. Another similar concern is absence of data that might result if a particular voting member of the committee remains silent or says little. Neither can strategic attempts by actors to hide preferences and thereby avoid revealing personal information be explicitly captured in this representation. Indeed, work by Pentland (2008) has shown that much social signaling occurs through body language and vocal dynamics that are not able to be captured in a transcript. It should, therefore, be clarified that this thesis does not claim that all social dynamics are manifest in language – rather, word-choice provides one source of insight into a complex, multi-modal process. The extent and severity of

this challenge is mitigated somewhat by the work of Boroditsky (2002, 2003), a cognitive scientist who has found evidence to indicate that not only does thought express itself through language, but that language use shapes patterns of thought. If such is the case, then differential use of language due, for example, to assigned roles, may reflect a salient role-based difference between decision-makers that is worth studying on its own merits (e.g., Simon 1964).

The approach presented here may be viewed as an extension of "latent coding" – one type of formal content analysis prevalent in the social sciences. The most important limitations of latent coding, and other hand-coding methods, are the inability to scale to large numbers of documents. This limitation stems from a dependence on the coder's knowledge, leading to inter-rater reliability concerns (Neuman 2005). Furthermore, hand-coding is labor-intensive, often requiring that teams of several coders be trained. The motivation behind using a computational approach is therefore to create a method that is automatic, repeatable and consistent. Quinn et al. (2006) provide a compelling justification for the adoption of computational text analysis techniques by social scientists. Furthermore, a computational method requires that the assumptions underlying the application of the methodology presented here are explicit, which enables a cumulative research paradigm. The work presented here therefore fits squarely in the tradition of statistical analysis of texts such as Network Text Analysis (Roberts 1997). Prominent examples in this tradition include cognitive mapping (Axelrod 1976) – a non-computational method for analyzing relations among causal structures, AutoMap (Carley 1992; Carley 1997; Diesner and Carley 2004) – a computational method currently under development, for extracting, analyzing and comparing mental models from texts based upon inferred and pre-defined conceptual categories, and Centering Resonance Analysis (Corman et al. 2002) – a method designed to identify and link important words within a discourse. Unlike these methods, in which the dominant paradigm is the representation of relations

among concepts, the ultimate goal of the method presented here is the analysis of relations among individual panel members.

Chapter Outline

We take the approach that a method which would accomplish the goals described above should aim for simplicity without being overly reductive. For example, one might simply count the number of each type of word that a particular speaker uses. This is a common approach favored by many users of latent coding methods. Due to the context-specific nature of the meetings analyzed, it is difficult to identify, *a priori,* words that might be important and thus, it is difficult to know which words to count or compare. We therefore begin our analysis with a method known as "Latent Semantic Analysis" (LSA), a natural language processing tool which was developed for purposes of information retrieval and topic grouping (Deerwester et al. 1990; Landauer et al. 1998). LSA was chosen because of its ability to identify a reduced number of putative concepts within a corpus. Two speakers who share similar concepts might therefore be related in some fashion. A preliminary study was performed with the goal of exploring the applicability of LSA to the problem space outlined above. Distributional assumptions underlying the application of LSA were found to introduce limitations that restricted its methodological applicability. These limitations were addressed using Latent Dirichlet Allocation (Blei, Ng, et al. 2003), a probabilistic model that circumvents many of the distributional assumptions underlying LSA. In particular, a variant of LDA, known as the Author-Topic (AT) model (Rosen-Zvi et al., 2004) was used, because of its ability to aggregate data from multiple speakers. For each transcript, the AT model was used to identify topics of interest to each speaker with the ultimate goal of constructing multiple probabilistic social affiliation network among topics. These networks were then aggregated to generate a representation of each meeting. Finally, for each meeting, temporal

information was incorporated. The final output is a directed graph, which may be interpreted as representing the flow of communication and influence within a given panel meeting.

Latent Semantic Analysis

One of the simplest computational approaches for analyzing terminology in context, is Latent Semantic Analysis (LSA) – a natural language processing tool which was developed for purposes of information retrieval and topic grouping (Deerwester et al. 1990; Landauer et al. 1998). LSA was initially created to address the issue of synonymy in information retrieval. Synonymy refers to the use of different words to represent the same concept (e.g., spice and seasoning). LSA addresses the synonymy issue through the use of Singular Value Decomposition (SVD), a technique from linear algebra that is aimed at determining a set of mutually orthogonal dimensions which describe the variance within a given set of data. When text data are analyzed, SVD tends to associate together words that have similar meanings. This is due to the empirical fact that words which have similar meanings tend to appear within the same contexts; i.e., words with similar meanings will co-occur either with each other or with the same sets of words. For example, one might encounter the following pair of sentences:

d_1: Pepper and salt add seasoning to the salad.
d_2: Pepper and salt are the two spices found most often in American restaurants.

These two sentences both contain the words "pepper" and "salt". From a brief overview of both documents, we would be able to infer that pepper and salt are seasonings (as in the first document), and that pepper and salt are spices. We would like to be able to infer that spices are seasonings.

The LSA Algorithm

Consider a corpus of documents, $\mathcal{D}$, containing n documents $d_1 \ldots d_n$. Consider, as well, the union of all words over all documents, W. Suppose there are $m>n$ words, $w_1 \ldots w_m$. We may therefore construct a "word-document matrix", **X**, with dimensions m x n, where each element in the matrix, x_{jk}, consists of a frequency count of the number of times word j appears in document k.

We conceive of the original word-document matrix as a "noisy" representation of word-word similarity (or document-document similarity). One source of this "noise" is the use of multiple words to represent the same concepts. We would therefore like to recover the original concepts implicit (or latent) in each word. Singular value decomposition with dimensionality reduction is a commonly used algorithm for the reduction of statistical noise. Using the above analogy, LSA performs noise reduction on the original word-document matrix.

The Singular Value Theorem in linear algebra states that any matrix may be represented as the product of three matrices, $\mathbf{X} = \mathbf{W}\,\mathbf{S}\,\mathbf{D}^T$, where **X** is the word-document matrix derived above. In this case, **W** is an $m \times m$ matrix of singular unit vectors, each of which are, by definition, mutually orthogonal. Each of these singular vectors corresponds to a word. Similarly, **D** is an $n \times n$ matrix of mutually orthogonal singular unit vectors. Each of the singular vectors in **D** corresponds to a document. Finally, **S** is an $m \times m$ diagonal matrix of decreasing, non-negative singular values, with each element corresponding to a linear combination of weights associated with each singular vector.

Without loss of generality, let r be the rank of **X**. In order to reduce the noise in **X**, we would like to reduce the rank of **X** such that $r' < r$ corresponds to the number of latent concepts within the corpus. We therefore set the smallest $(r-r')$ singular values to 0, generating **S'**. The value of r' must be chosen by the user, although values of r' between 100-300 seem to work well for information retrieval

purposes (Landauer et al. 1998). The resulting matrix, $\mathbf{X' = W S' D^T}$, is a rank *r'* approximation of **X** that can be represented as having *r'* mutually orthogonal singular vectors. Words and documents, which were previously represented by linear combinations of *r* mutually orthogonal singular vectors, are now represented as linear combinations of *r'* mutually orthogonal vectors, such that the locations of words and documents in the vector space represented by **X'** approximate the corresponding locations of words and documents in **X** in a least-squares sense. If we were to treat **X'** as a Euclidean space, the normalized inner product of (i.e., the cosine between) two word-vectors (represented as rows of the matrix **W S'**) can be thought of as the projection of each word upon a set of axes, each of which corresponds to a latent concept. Therefore, this value would correspond to the two words' degree of synonymy (or similarity for documents). LSA is therefore able to capture higher-order relations between synonymous words (e.g., words that do not directly co-occur, but that mutually co-occur with a third word as in the spice/seasoning example above).

LSA Implementation

LSA was implemented in Python 2.5 and MATLAB. Python 2.5 was used to parse an FDA Advisory Panel meeting into a word-document matrix, which was then imported into MATLAB. One possible source of measurement error includes the existence of various forms of word conjugation (e.g., patient vs. patients) that might be classified as different words. In general, syntactic information is not captured by the "bag-of-words" representations of corpora used in this thesis. The use of stemmer algorithms (such as PyStemmer[2]) are aimed at eliminating some of this error. Finally, frequently-occurring, but non-content-bearing words (such as "the", "and", "a", etc.,) can skew results. Error due to this problem is eliminated through the incorporation of a "stop list", which automatically removes these words. The stop list used for the analyses in

[2] Available at: http://sourceforge.net/projects/pystemmer/

this thesis was compiled by the Semantic Indexing Project (http://www.knowledgesearch.org) and is shown in Appendix 1. In the LSA approach, remaining error associated with non-content-bearing words is managed by the use of log-entropy weighting[3] (Dumais 1991). Singular value decomposition and log-entropy weighting were executed using built-in MATLAB functions, generating an LSA space. Finally specialized functions were written to perform the coherence analyses described below. These approaches are typical in natural language processing (Manning and Schütze 1999).

Other applications of LSA have included automated student essay evaluation (Landauer and Dumais 1997), measurement of textual coherence (Foltz et al. 1998), knowledge assessment (Rehder et al. 1998), information visualization (Landauer, Laham et al. 2004), the quantitative analysis of design team discourses (Dong et al. 2004), and the construction of a theory of human learning and cognition (Landauer and Dumais 1997). In particular, Dong (2005) has used LSA to study conceptual coherence in design and the process by which members of a design team agree upon a common design representation. This work begins by extending Dong's techniques to the realm of advisory committee decision-making, and is ultimately meant to contribute a data-driven methodology that may provide insight into the effects of institutional background on decision-making for complex engineered devices and systems.

Committee Textual Coherence as a Metric

The use of LSA to measure textual coherence can provide insight into the extent to which different speakers within an advisory panel meeting are using

[3] Log-entropy weighting is applied to a word-document matrix in order to improve its information-retrieval performance. This has the effect of emphasizing words that are unique to a given speaker, thereby enabling a focus on his/her unique language characteristics. The formula consists of two coefficients: the term weight, t, is simply $\log(1+f)$, where f is the frequency of a specific word in a given document; the global weight, g, is $p*\log(p)$, where p is the ratio of times that a specific word appears in a given document to the number of times that word appears in all documents. The log-entropy weight is simply $t*g$.

terminology in the same way. Coherence analysis was first implemented in (Foltz et al. 1998), and extended to design teams in (Dong 2005). In a design team, designers must be "on the same page". This means that they must be speaking in words that are sufficiently similar as to be comprehensible to each other, i.e., speaking similar professional languages. LSA does allow for the analysis of relative linguistic homogeneity, thereby enabling a determination of the extent to which designers are "on the same page" relative to one another through a coherence metric.

Medical advisory panels may be equivalently viewed as teams (McMullin and Whitford 2007). Although they are not designing an artifact, as in Dong's work, such panels must produce a policy recommendation that will have a strong impact upon the success or failure of the technical system under review and thus have an overall common objective. Our approach is to use LSA to study mutual understanding within medical advisory panels by studying the respective coherence of one actor as compared to another. Given that committee members vote, voting records provide a measurable source of data against which to compare LSA performance, a test not available to Dong in his studies of design teams.

Preliminary Results from LSA

Shown below are the results from a preliminary analysis of a meeting of the Circulatory Systems Devices Advisory Panel Meeting held on April 22, 2005 using LSA. In this panel meeting, the Circulatory System Devices Advisory Panel discussed and made recommendations regarding the approval of the "PAS-port", a device aimed at reducing the risk of stroke inherent in coronary artery bypass (Maisel 2005). This device was under review for 510(k) approval when its predicate device was pulled from the market. This had the effect of prompting the FDA to create new requirements for similar devices. Since the predicate

device was now invalid, the PAS-Port device was brought to the advisory for review despite the fact that the sponsors had initially not planned to execute full-fledged clinical trials. The device's sponsors used observational data from two clinical trials conducted outside of the United States, and therefore, under different conditions than those which might have been required by the FDA had they been conducted under an Investigational Device Exemption (IDE) for a PMA. As a result, there were several questions regarding the viability of the data (and hence, the sponsor's contention that the device was safe). Among these were the following:

1. The sponsor's presentation attempted to combine the results of two clinical trials conducted under different conditions. Thus, there was a question of whether the data could be pooled to yield meaningful results.

2. Following the failure of the predicate device, the FDA increased the lower bound for the confidence interval surrounding a proposed device's patency rate (i.e., the rate at which a vein graft would remain un-blocked). This implied that a statistical test with higher power was required. Nevertheless, these new requirements occurred after the sponsor had already run the clinical trials.

3. The data were collected outside of the United States, and therefore, were not supervised by the FDA. Rather, the studies were designed for European clinical trial reviewers.

4. The device under study was improved between clinical trials, thereby leveraging the experience of the designers to improve its safety and efficacy, but simultaneously contributing to the non-comparability of the two trials.

Examination of Top Five Log-Entropy Words

In order to determine whether different actors do use substantively different terminology, we examined the top five log-entropy-weighted stemmed words for each speaker. Table 1 demonstrates the results of this analysis. A qualitative analysis of this table shows that different speakers' unique terminology relates to their roles in the meeting. For example, the chairman seems to use largely procedural rules, the executive secretary is using words associated with conflict of interest regulations, etc.

Table 1: Listing of the top five log-entropy weighted words for each speaker.

Speaker Occupation	Word 1	Word 2	Word 3	Word 4	Word 5
Chairman	'jeff'	'move'	'norm'	'session'	'afternoon'
Executive Secretary	'waiver'	'compet'	'firm'	'conflict'	'particip'
FDA Representative	'landscap'	'stori'	'krucoff'	'stratifi'	'agenc'
Cardiologist	'late'	'surrog'	'inpati'	'prevent'	'fitzgibbon'
Cardiologist	'overt'	'iter'	'concept'	'flesh'	'engin'
Cardiologist	'ultim'	'lumenolog'	'behavior'	'concord'	'behav'
Statistician	'variabl'	'henc'	'certainti'	'school'	'weight'
Cardiac Surgeon	'censor'	'cleveland'	'wider'	'precious'	'obliqu'
Cardiologist	'draw'	'extrapol'	'feasibl'	'popul'	'electrocardiogram'
Cardiac Surgeon	'room'	'vote'	'forth'	'esteem'	'variabl'
Pharmacologist	'gray'	'75'	'zone'	'noncompar'	'variat'
Cardiac Surgeon	'handl'	'stroke'	'calcifi'	'unfortun'	'val'
Cardiologist	'cath'	'old'	'anybodi'	'struck'	'catheter'
Cardiac Surgeon	'shower'	'disturb'	'biggest'	'hole'	'clamp'
Industry Representative	'agenc'	'salvag'	'therapeut'	'vice'	'proxima'
FDA	'feedback'	'track'	'premarket'	'piec'	'cdrh'
FDA	'gore'	'approv'	'program'	'recommend'	'pma'
Sponsor CEO	'endotheli'	'japan'	'surfac'	'amount'	'tool'
Sponsor Clincian	'remain'	'convert'	'literatur'	'sequenti'	'stenosi'
Sponsor Statistician	'adjust'	'valu'	'strata'	'26'	'bucket'
Sponsor Surgeon	'connector'	'saphen'	'sutur'	'spot'	'endoscop'

FDA	'preclin'	'element'	'stainless'	'implant'	'flang'
FDA	'pivot'	'covari'	'visit'	'intraop'	'itt'
FDA	'undertaken'	'recruit'	'sutur'	'input'	'pivot'
FDA	'side'	'ophthalm'	'yue'	'lili'	'variabl'
Consumer Representative	'decreas'	'cool'	'aw'	'61'	'deployt'

Cluster Analyses

Prior to the execution of the Latent Semantic Analysis, individuals' utterances were grouped together into speaker-specific vectors. This was accomplished by adding together the vectors for each of their utterances. A k-means clustering algorithm was then run on these vectors in an attempt to separate the actors into two clusters. This generated clusters that corresponded to advisory panel members and FDA or sponsor representatives. Table 2 outlines the results:

Table 2: "Confusion Matrix" for stakeholder cluster analysis. ($p = 9.97 \times 10^{-5}$; $\chi^2 = 15.14$; $df = 1$)

	FDA or sponsor reps.	Panel Members
Cluster 1	2	14
Cluster 2	9	1

Some meaning might be imputed to these clusters. Cluster 1 is largely made up of the Panel Members who can be thought of as "non-partisans" whereas cluster 2 largely consists of potentially "partisan" FDA or sponsor representatives. The "non-partisan" who was incorrectly classified corresponds to the panel's executive secretary, whose primary role at the end of the meeting was to read questions posed by the FDA. On the other hand, the two "partisans" who were incorrectly classified were the sponsor's statistician, who interacted directly with the panel, and the FDA representative to the advisory panel, who serves a dual role as panel member and whose job it is to oversee the advisory panel proceedings.

Attempts to separate actors into more clusters resulted in other subdivisions of the group which could be explained in terms of combinations of formal roles, training (e.g., statisticians, cardiologists, etc.), partisanship, frequency of speech, and random assignment due to noise. Often, a clustering algorithm would yield a small number (5 or less) of clusters roughly corresponding to identifiable groups, with most other clusters assigned randomly or by frequency of speech. Therefore, the ability of the cluster analysis to reliably perform fine divisions among panel members is questionable.

Coherence Analysis

Following (Dong 2005), an LSA-based coherence analysis of the meeting was performed. Figure 4 shows an analysis of the meeting described above. For the purposes of this analysis, actors were categorized into four bins: Voting members; FDA; Sponsors; and Non-Voting Members. Each time series represents the running average of the semantic coherence of a particular group, measured with respect to the final semantic coherence of the voting members. Running average semantic coherence, $c(\tau) = \cos(\theta)$, where θ is the angle between two vectors $\mathbf{s}(\tau)$, the running average centroid of speaker s at time τ, and $\mathbf{v}$, the centroid of the voting members at the end of the discourse. $\mathbf{s}(\tau) = \frac{\sum_{t=0}^{\tau} u(t)}{n(\tau)}$, where $\mathbf{u}(t)$ is the location of utterance t in the semantic space, and $\mathbf{n}(\tau)$ is the number of utterances spoken by speaker s at time τ. In this case, each "speaker" is actually a group of speakers, constituting the FDA representatives, the voting members, the non-voting members, and the sponsors. The sponsor's coherence with respect to the voting members' final position (the dashed line) drops dramatically around utterance number 200, hitting its minimum at utterance number 218, as indicated by the large, negative slope for the sponsor's coherence time-series curve. Note that the FDA's coherence (the dashed and dotted line) also drops with respect to

the voting members. This is likely due to the fact that the voting members did not focus on all of the FDA representatives' arguments.

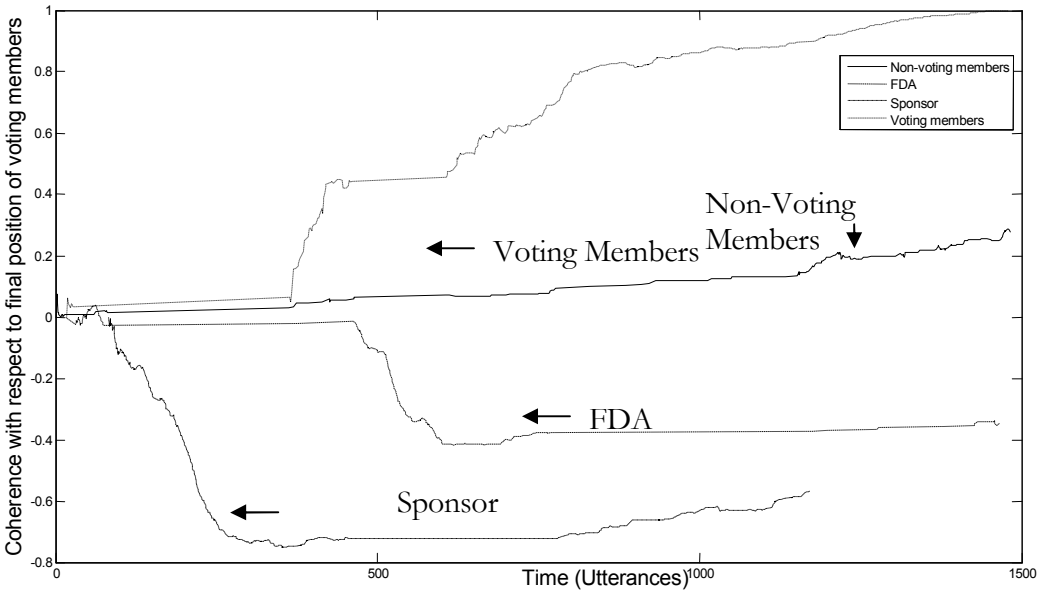


Figure 4: Coherence of group centroid with respect to final centroid of voting members. The horizontal axis, representing the utterance number in the discourse, represents progress through the discourse. The vertical axis is coherence as measured with respect to the final position of the voting members. Each curve corresponds to a different group present at the meeting (Non-voting panel members, FDA representatives, sponsors, and voting members).

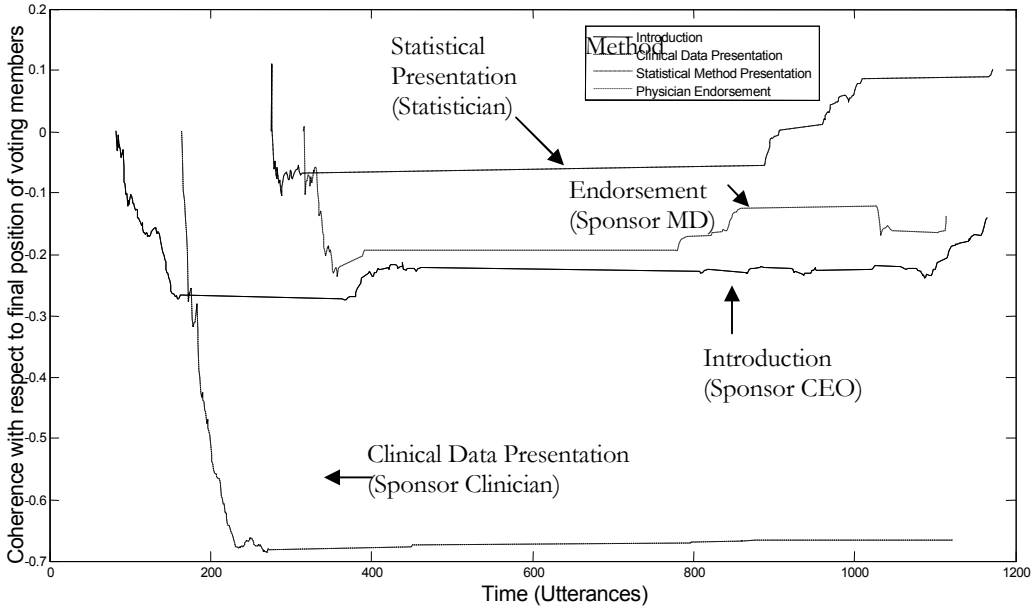


Figure 5: Breakdown of sponsor's presentation by speaker/section. Each curve corresponds to a different speaker, each of whom presented a different phase of the sponsor's talks (introduction, clinical data, statistical method, and physician endorsement).

Figure 5 further examines the sponsor's coherence measured with respect to the final position of the voting members in semantic space. The greatest drop occurs at the time of the presentation of the data, at utterance 218. This is consistent with a focus on the data that is different from that used by the panel members. Although there was some disagreement regarding the viability of pooling the two clinical trials together (captured in the Statistical Methods Presentation), most of the later discussion focused on the interpretation of the data rather than on the methods used to reach that interpretation.

Analyses of other meetings yield results that are qualitatively very similar, suggesting that the LSA approach outlined above is not actually extracting information that is specific to each meeting – rather, it is reflecting the procedural aspects of the meeting. In particular, it stands to reason that the sponsors and the FDA would use very different language than the panel members would, simply by virtue of their role. Hence, this analysis requires a tool that is more sensitive to individual differences between speakers. As an introduction, an overview of some of the limitations of the LSA approach is therefore essential.

Limitations of the LSA approach

Construction of LSA Metrics

Differences between Dong's and Foltz's approaches in studying coherence highlight some of the limitations of using LSA for representing coherence in design teams. Whereas Foltz studied individual students writing essays about a well-defined topic, Dong studied interactions between multiple people attempting to discuss a design that has not yet been produced. Therefore, Foltz pre-trained LSA on domain-relevant materials, whereas Dong explicitly did not do so because it would bias the outcome in favor of a particular design or method (Dong 2007, personal communication). Group coherence, as defined above, is a relative measure.

Representing group coherence over the course of a discussion is challenging. Because coherence is defined as a distance metric between two utterances, there is no natural baseline against which to evaluate the coherence of a particular statement. Dong uses the group document Euclidean centroid as this baseline. This has suspect validity because the group document centroid is not necessarily representative of the instantaneous coherence between two adjacent utterances at any point in time. Indeed, as the overall coherence of the design discussion decreases, we may consider the document centroid to be an increasingly worse

representation of the group's "shared mental model", presuming one exists. An assumption that group members share a mental model stands in contrast to an analysis of differences in perspectives on committees of technical experts. Furthermore, the use of a running average metric has a tendency to overemphasize early statements and damp out later ones because early points in the time series carry more weight when compared directly with the group document centroid than do later points. Later points are averaged with all of the earlier points and therefore may lose important dynamics. This leads to a result that always converges, by definition, to unity. This can impart a significant bias on the time series results since later statements will tend to have a higher coherence than do earlier ones, giving the potentially false impression that the conversation is converging when it may not be. The comparison of individuals' utterances against the group centroid shows how an individual may be converging to the ultimate group decision, although it provides little information about how individuals interact with one another.

Choice of Document Size

There are many different ways of dividing a discourse, each of which might yield slightly different results. Use of a court-recorder's discretion in defining the boundaries of an utterance typically ensures conceptual coherence, at the cost of a potential source of subjectivity. Furthermore, very short utterances might make less of a contribution to the analysis than do longer utterances. An ideal approach would incorporate both temporal information and authorship information sources. The former would enable some insight into meeting dynamics, whereas the latter would allow for sufficient data to enable a statistically meaningful characterization of a given speaker's position relative to others.

Dimensionality Selection

There is currently no theoretically optimal number of dimensions for a latent semantic analysis. Foltz and Dong used different dimensionality reduction choices when calculating their respective LSA spaces. Foltz, possessing a training set against which to measure, kept 300 latent dimensions (i.e., $r' = 300$) using LSA as described above. Dong, on the other hand, kept dimensions 2-101, a technique first used in (Hill et al. 2002) that removes the largest-weighted singular value. This seems to under-emphasize effects due to word frequency and direct word co-occurrence, and over-emphasize higher-order co-occurrence. This difference highlights the fact that there is currently no theoretical optimum for determining the appropriate number of dimensions in an LSA analysis. Attempts to calculate such an optimum suggest that it may be highly dependent on the structure of the individual discourse and the nature of the query (Dupret 2003).

Polysemy

Although LSA is largely successful in reducing the synonymy problem by grouping words together that appear in the same context, the polysemy problem – encountered when two words have the same spelling but different meanings (compare "bat" the animal vs. "bat" in the context of baseball) – is not well addressed by this specific methodology. This is because polysemous words are typically represented as weighted averages between any number of document vectors. Rather than being assigned multiple times to different meanings/contexts, polysemous words are represented inaccurately as the weighted average of those contexts. Among these is the assumption that words are embedded within a Euclidean "semantic-space". This particular assumption breaks down when comparing words that are polysemous. LSA represents the location of these words in the Euclidean semantic space as the average over the two separate locations – an incorrect representation. LSA is known to be vulnerable to problems of polysemy. As such, use of LSA to analyze

conversations is likely to be susceptible to this weakness. Medical device approval committee meetings may tend to avoid polysemy because of the use of highly specialized and well-defined professional terms by the physicians, statisticians and health policy experts performing the evaluation. On the other hand, those words that are most likely to be ambiguously defined, and thus most interesting, are *de facto* polysems. Such words as "safety" and "efficacy", whose meanings must be *defined* relative to a device during these FDA meetings are likely to be sources of debate. This is a major limitation of LSA, that has been overcome by existing algorithms designed to solve problems associated with polysemy (Blei et al. 2003; Dhillon and Modha 2001; Hofmann 2001).

Unrealistic modeling assumptions

Papadimitriou et al. (2002) explain the empirical success of LSA by formulating it as a probabilistic model. In the process, the authors make explicit the statistical distribution assumptions that underlie the LSA approach. LSA assumes linearity for a set of latent dimensions underlying a Euclidean semantic space. Furthermore, each word's location in the Euclidean space is linearly-distributed, an assumption that introduces increasingly more distortion into the analysis as a given speaker uses fewer words. These limitations make it difficult to resolve the linguistic attributes of individual speakers, particularly in the absence of extensive speaker data within a given meeting. Furthermore, the latent dimensions of the LSA feature space, which nominally correspond to latent concepts of a discourse, are often difficult to interpret.

Bayesian Topic Models

The leading alternative to LSA is Latent Dirichlet Allocation (LDA), a Bayesian "topic model" (Blei et al. 2003). For an excellent comparison of LSA to Bayesian models of text analysis, see (Griffiths et al. 2007). Approaches based on Bayesian inference, such as Latent Dirichlet Allocation (LDA), provide a platform that may

be used to avoid many of the limitations noted above. Of particular interest are topic-modeling approaches to studying social phenomena in various contexts. Topic models have been applied to the social sciences in a limited fashion, with examples having largely taken the form of studies of the evolution of specialized corpora (Hall et al. 2008), analysis of structure in scientific journals (Griffiths and Steyvers 2004), finding author trends over time in scientific journals (Rosen-Zvi et al., 2004), topic and role discovery in the Enron email networks (McCallum et al. 2007), analysis of historical structure in newspaper archives (Newman and Block 2006), and group discovery in socio-metric data (Wang et al., 2005). Topic models have also been applied to other fields, using, for example, genomic data as input.

Topic Models address the limitations of LSA

Unlike LSA, which uses a continuous Euclidean metric space representation, LDA assumes probabilistic assignment of each word to a discrete topic. Each topic is assumed to be *exchangeable*, i.e., conditionally independent of each other topic (de Finetti 1974). LSA's assumption of orthogonal latent dimensions in a Euclidean space implies that each word can be located by a unique point in that semantic space. LDA's exchangeability assumption, on the other hand, allows for words to have multiple "senses" – i.e., the same word may occur in two different topics. Rather than modeling a word as an average between two locations in a latent Euclidean space, a word is instead modeled as having been drawn from a discrete probability distribution over topics. This provides a natural solution to the polysemy problem (Griffiths et al. 2007). The basic structure of an LDA model is shown in Figure 6.

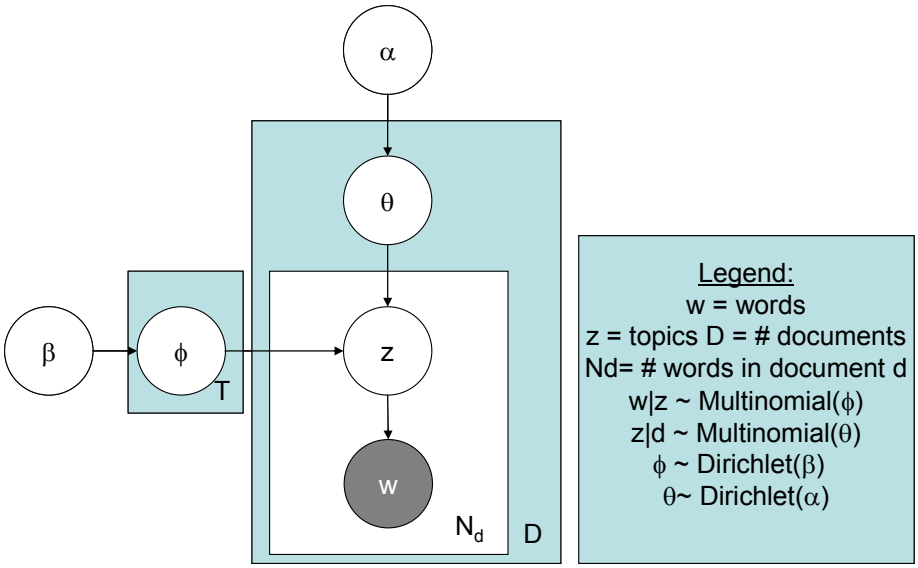


Figure 6: A plate-notation representation of the Latent Dirichlet Analysis algorithm (Blei, Ng, et al. 2003).

In an LDA model, words are observed in a word-document matrix, as in LSA. Each word (w) is assumed to be drawn from a *topic* (z). A topic is accordingly defined as a multinomial distribution (ϕ) over words (i.e., a word is chosen at random by rolling a weighted w-sided die, where w is the total number of words in the corpus). Each document is similarly modeled as a multinomial distribution (θ) over topics. The parameters (i.e., the die-weights) for each multinomial distribution are drawn from a symmetric Dirichlet prior distribution – a multivariate distribution that is the conjugate prior to the multinomial distribution. Each Dirichlet distribution has a number of parameters equal to the number of parameters of its corresponding multinomial distribution. Nevertheless, early LDA models all assume that the Dirichlet priors are symmetric – i.e., all of the parameters are the same. The utility of this assumption has only recently been tested (Wallach and McCallum 2009). The "hyperparameters" defining each Dirichlet prior (α and β) are chosen by the

modeler, and are the primary means, along with choosing the number of topics, by which the form of the model might be controlled. These hyperparameters may be interpreted as smoothing parameters. In particular, if $0 < \alpha < 1$, topics are very document-specific, whereas for values of $\alpha > 1$, topics are smoothed across documents. Similarly, if $0 < \beta < 1$, words are very specific to topics (i.e., there is relatively little polysemy), whereas for values of $\beta > 1$, words are smoothed across topics. Thus, topic models may account for polysemy in a way that LSA cannot.

LDA defines a family of probabilistic models, and must be fit to a specific corpus using Bayesian inference algorithms. We are interested in finding the most probable hypothesis, h, (i.e., the most appropriate model), given the observed data, d. Model fitting may be explained using Bayes' theorem:

$$p(h \mid d) = \frac{p(d \mid h) * p(h)}{p(d)} \quad \textbf{(1)}$$

Using the notation specific to the LDA model, Bayes' theorem may be expressed as follows (Blei et al. 2003):

$$p(\theta, z \mid w, \alpha, \beta) = \frac{p(\theta, z, w \mid \alpha, \beta)}{p(w \mid \alpha, \beta)} \quad \textbf{(2)}$$

This guarantees that the topics that are inferred by LDA are appropriate to the corpus being analyzed. Furthermore, the presence of the Bayesian priors ensures that the model is not over-fit to the corpus data – a limitation that had been encountered by previous attempts to generate a probabilistic form of LSA (Hofmann 2001). The flexibility of LDA's representation therefore circumvents the problems encountered by LSA as a result of its distributional assumptions. This comes at the cost of computational efficiency, since explicit computation of LDA's posterior distribution (i.e., the distribution that we would like to determine in order to be able to fit topics to the data) is intractable. To see why this is, we

must expand the expression above into its constituent parts. The numerator is easily expanded using the canonical expressions for the multinomial and Dirichlet distributions.

$$p(\theta, z, w \mid \alpha, \beta) = p(w \mid \phi) * p(\phi \mid \beta) * p(z \mid \theta) * p(\theta \mid \alpha)$$
$$= C_1(\alpha, \beta) * (\prod_{i=1}^{V} \phi_i^{w_i}) * (\prod_{j=1}^{V} \phi_j^{\beta-1}) * (\prod_{k=1}^{T} \theta_t^{z_t}) * (\prod_{k=1}^{T} \theta_t^{\alpha-1}) \quad \textbf{(3)}$$

Here, V is the total number of words in the corpus and T is the total number of topics. C_1 is a term whose value is a function only of the hyperparameter values. It serves as a normalizing parameter. As can be seen from this expression, the Dirichlet distribution may be interpreted as a "virtual count" – i.e., the hyperparameters may be interpreted as presumed data that has already been seen and added to the observed data. The Dirichlet prior therefore may be said to reflect one's prior beliefs regarding the propensity of a particular topic or word in the data. The denominator is not analytically tractable, and may be expressed as follows (Blei et al. 2003):

$$p(w \mid \alpha, \beta) = C_2(\alpha) * \int (\prod_{i=1}^{T} \theta_i^{\alpha-1}) * (\prod_{n=1}^{D} \sum_{j=1}^{T} \prod_{k=1}^{V} (\theta_j \beta_{jk})^{w_n^k}) d\theta \quad \textbf{(4)}$$

The original implementation of LDA inferred the posterior distribution for its test corpus using a technique known as variational inference (Blei et al. 2003). Nevertheless, variational inference has not been widely adopted by the topic modeling community because it is difficult to implement and because it lacks a theoretical guarantee that it will converge to the global maximum of the posterior distribution. Gibbs sampling, a Markov-Chain Monte Carlo (MCMC) method, adopted from statistical physics, possesses such a guarantee and, although

potentially slower[4], is currently in widespread use among topic modelers (Griffiths and Steyvers 2004). Gibbs sampling for LDA proceeds following Algorithm 1:

Algorithm 1: LDA Implementation

1. Initialize topic assignments randomly for all word tokens
2. **repeat**
3. **for** d=1 to D **do**
4. **for** i=1 to N_d **do**
5. draw z_{di} from $P(z_{di} | \mathbf{z}_{-di}, \mathbf{w}, \alpha, \beta)$
6. assign z_{di} and update count vectors
7. **end for**
8. **end for**
9. **until** Markov chain reaches equilibrium

Here, D is the total number of documents and N_d is the number of word tokens in each document, z_{di} is the topic assigned to word token i in document d. The non-normalized form of $P(z_{di} | \mathbf{z}_{-di}, \mathbf{w}, \alpha, \beta)$ is derived in (Griffiths and Steyvers 2004) as follows:

$$P(z_{di} = j \mid z_{-di}, w, \alpha, \beta) \propto \frac{n_{-di,j}^{(w_{di})} + \beta}{n_{-di,j}^{(\cdot)} + V\beta} \frac{n_{-di,j}^{(d)} + \alpha}{n_{-di,j}^{(\cdot)} + T\alpha} \quad \textbf{(5)}$$

[4] Recent work by (Goodman, Mansinghka et al. 2008) has focused on generating a probabilistic programming language whose purpose is to enable fast Bayesian inference of the sort required for these analyses. Such research could significantly increase the adoption of MCMC-type algorithms as they become more easy, and faster, to implement. This could increase the rate of adoption of MCMC over variational inference even further.

In this expression, $n^{(a)}_{-b,c}$ is a count vector – i.e., a count of the number of times all tokens with identity a (e.g., all words with identity w_i or all tokens in document d), excluding token b are assigned to topic c. $n^{(\cdot)}_{-b,c}$ denotes that all tokens assigned to topic c should be considered, regardless of word or document identity, with the exception of token b. V is the total number of unique words in the corpus, and T is the total number of topics. As can be seen from the form of the above expression, each token's probability of being assigned to a given topic is proportional to the number of times that that word appears in that same topic, and to the number of times a word from that document is assigned to that topic. This defines a Markov chain, whose probability of being in given state is guaranteed to converge to the posterior distribution of the LDA model as fit to the corpus after a sufficiently large number of iterations. Evaluating Markov chain convergence is currently an open area of research. It is therefore standard practice for a Markov chain to be run for multiple iterations in order to ensure convergence. These initial iterations are known as a "burn-in" period. Throughout this thesis, burn-in length is set to 1000 iterations – a value frequently used in the topic modeling literature (Griffiths and Steyvers 2004).

The ability to fit the probability distribution underlying the LDA model to a specific corpus neatly solves the problem inherent in the statistical distribution assumption underlying LSA. Once the LDA model has been defined, variants may be utilized, given the nature of the problem being solved. In particular, the LDA model as outlined above is still sensitive to the arbitrary document boundaries imposed by the court recorder. Furthermore, documents vary significantly in length – some might only be two words (e.g., "Thank you") whereas others might be significant monologues. One possible approach would be to attempt to group documents together by their temporal ordering – this would enable groups of documents to pool statistical power. This is the approach taken by the work of (Quinn et al. 2006), who built a topic model of US Senate

proceedings with the goal of identifying agenda-setting behavior. Quinn's model explicitly incorporated time as an explanatory variable, using techniques pioneered by (Pole et al. 1994; Blei and Lafferty 2006). In practice, this approach serves to smooth topics across time such that they rise and fall in a continuous fashion. The temporal element of the analysis enables insight into how agendas in the US Senate are built and changed. (Fader et al. 2007) used the resulting dataset to identify influential members of the US Senate using a technique knows as "MavenRank". Influence was operationalized as "lexical centrality" – a metric of similarity between a given speaker's utterances and all other utterances by speakers using that topic. (Fader et al. 2007) present results indicating that lexical centrality is associated high-status positions within Senate committees. Preliminary tests of MavenRank on FDA panels suggests that an individual's influence under the lexical centrality scheme is strongly correlated with their air-time in that topic (i.e., the number of utterances that they express). Findings in social psychology dispute the relation between actual influence and air-time (Bottger 1984) suggesting that MavenRank captures procedural sources of influence instead of actual influence. Furthermore, MavenRank is unable to determine influence *across* topics, and therefore cannot address questions regarding why one topic might come to prominence within a given committee meeting. Indeed, Fader et al. treated specific topics as standing in fixed association with specific committees – an assumption that makes sense given the standing committees in the US Senate, but is inappropriate for the more flexible committees to be found in FDA and other engineering systems. Furthermore, in applying their technique, Quinn et al. fit a model to several different simultaneous Senate discourses. Data on FDA panels is much more linear, since it represents one conversation rather than several years' worth of speeches[5]. Therefore the

[5] On the importance of linearity in speech, see (Gibson 2005).

approach of Quinn et al. is not directly applicable to the problem addressed in this thesis.

The Author-Topic Model

A variant of LDA, the Author-Topic (AT) model, can been used to generate a distribution over topics for each participant in a meeting (Rosen-Zvi et al. 2004). Given that the literature suggests that each speaker possesses an institutional or role-based signature in his or her word choice, we would like to have the identity of the speaker inform the selection of topics. We therefore use a variant of Rosen-Zvi et al.'s Author-Topic (AT) Model (2004), which creates probabilistic pressure to assign each author to a specific topic. Shared topics are therefore more likely to represent common jargon. The Author-Topic model provides an analysis that is guided by the authorship data of the documents in the corpus, in addition to the word co-occurrence data used by LSA and LDA. Each author (in this case, a speaker in the discourse) is modeled as a multinomial distribution over a fixed number of topics that is pre-set by the modeler. Each topic is, in turn modeled as a multinomial distribution over words. A plate-notation representation of the generative process underlying the Author-Topic model is found in Figure 7.

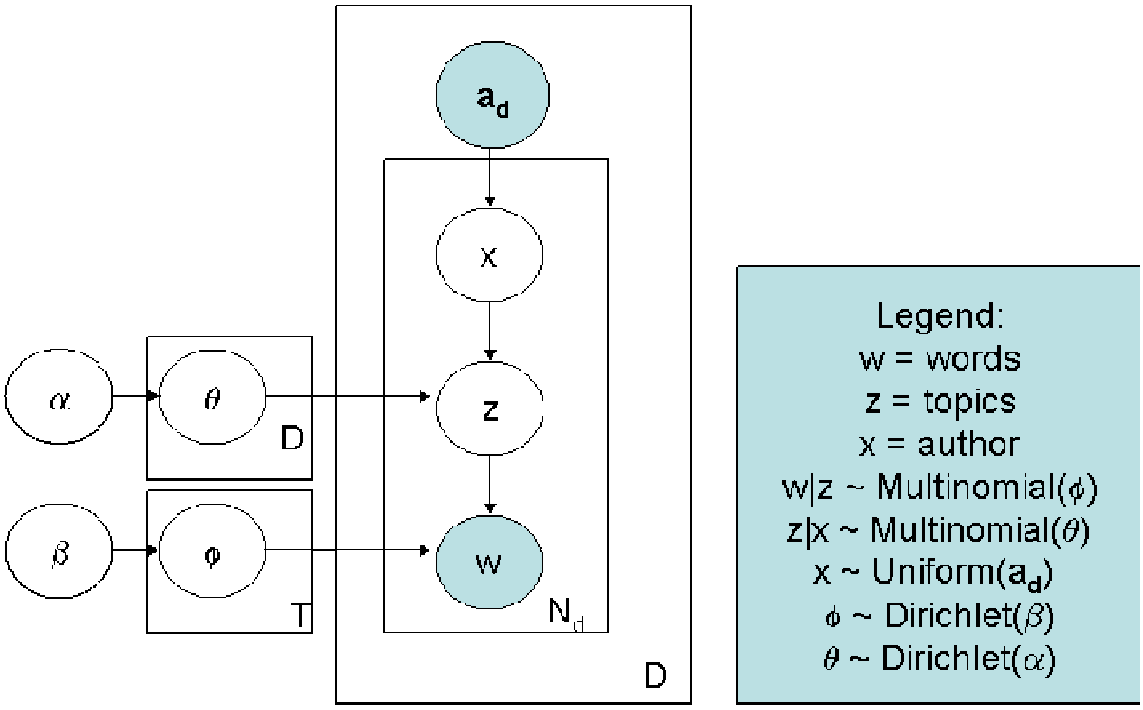


Figure 7: A plate notation representation of the Author-Topic model from (Rosen-Zvi et al. 2004). Authors are represented by a multinomial distribution over topics, which are in turn represented by a multinomial distribution over all words in the corpus.

As an LDA variant, the Author-Topic model is populated using a Markov-Chain Monte Carlo Algorithm that is designed to converge to the distribution of words over topics and authors that best matches the data. Information about individual authors is included in the Bayesian inference mechanism, such that each word is assigned to a topic in proportion to the number of words by that author already in that topic, and in proportion to the number of times that specific word appears in that topic. Thus, if two authors use the same word in two different senses, the AT Model will account for this polysemy. Details of the MCMC algorithm

derivation are given in (Rosen-Zvi et al. 2004). The AT model was implemented in MATLAB by the author, based on the Topic Modeling Toolbox algorithm (Griffiths and Steyvers 2004). Gibbs sampling for AT proceeds following the algorithm in Algorithm 2:

Algorithm 2: AT Model Implementation Algorithm

1. Initialize topic assignments randomly for all word tokens
2. **repeat**
3. **for** d=1 to D **do**
4. **for** i=1 to N_d **do**
5. draw z_{di} & x_{di} from $P(x_{di},z_{di} | \mathbf{z}_{-di}, \mathbf{x}_{-di}, \mathbf{w}, \alpha, \beta)$
6. assign z_{di} & x_{di} and update count vectors
7. **end for**
8. **end for**
9. **until** Markov chain reaches equilibrium

Here, D is the total number of documents and N_d is the number of word tokens in each document, z_{di} is the topic assigned to word token i in document d, and x_{di} is the author assigned to word token i in document d. The form of $P(x_{di},z_{di} | \mathbf{z}_{-di}, \mathbf{x}_{-di}, \mathbf{w}, \alpha, \beta)$ is derived in (Rosen-Zvi et al. 2004) as follows:

$$P(z_{di} = j, x_{di} = k \mid z_{-di}, x_{-di} w, \alpha, \beta) \propto \frac{n_{-di,j}^{(w_{di})} + \beta}{n_{-di,j}^{(\cdot)} + V\beta} \frac{n_{-di,j}^{(x_{di})} + \alpha}{n_{-di,j}^{(\cdot)} + T\alpha} \quad \textbf{(6)}$$

As can be seen from the form of the above expression, each token's probability of assignment to a given topic is proportional to the number of times that that word appears in that same topic, and to the number of times a word from that

author is assigned to that topic. This defines a Markov chain, whose probability of being in given state is guaranteed to converge to the posterior distribution of the AT model as fit to the corpus after a sufficiently large number of iterations.

Under the special case where each document has one unique author, the AT model is equivalent to LDA. Similarly, under the special case where each document has one non-unique author, the AT model is equivalent to an LDA model where each author may be treated as one document. As will be shown below, the multiple authorship feature of the AT model may be used to determine a given speaker's idiosyncratic word choice, a useful feature when many panel members share procedural language that may not necessarily be related to their decisions.

Hyperparameter Selection

Like LDA, the AT model requires the selection of two parameters. Ideally, we would like to determine which parameters used for the AT model best fit the corpus data. We must simultaneously be wary of over-constraining the analysis with the assumptions underlying the AT model. A popular metric for goodness-of-fit used within the machine learning literature is *cross-entropy* (Manning and Schütze 1999). Cross-entropy is a metric of the average number of bits required to describe the position of each word in the corpus and is closely related to *perplexity*. Both perplexity and cross-entropy are closely related to *log-likelihood*, a measure of how well a given model predicts a given corpus. Therefore, lower perplexity indicates a more parsimonious model fit. Lower perplexity also indicates that the assumptions underlying the model are descriptive of the data. For the AT model, cross-entropy may be calculated as follows:

$$H(p,q) = -\frac{\sum_{i=1}^{N} \log_2 (p(w_i \mid \phi, \beta) * p(z_i \mid \theta, \alpha) * p(x_i))}{N} \quad \textbf{(7)}$$

In this expression, N is the total number of word tokens. The expression in the numerator is the empirical *log-likelihood* (although log-likelihood is usually calculated using a natural logarithm). Thus, a natural interpretation of cross-entropy is the average log-likelihood across all observed word-tokens. Perplexity is defined as $2^{H(p,q)}$. The lower a given model's perplexity or cross-entropy, or the higher its log-likelihood, the more parsimonious is the model's fit to the data.

Each author's topic distribution is modeled as having been drawn from a symmetric Dirichlet distribution, with parameter α. Values of α that are smaller than unity will tend to more closely fit the author-specific topic distribution to observed data – if α is too small, one runs the risk of overfitting. Similarly, values of α greater than unity tend to bring author-specific topic distributions closer to uniformity. A value of α=50/(# topics) was used for the results presented in this thesis, based upon the values suggested by (Griffiths and Steyvers 2004). For the numbers of topics considered in these analyses (generally less than 30), this corresponds to a mild smoothing across authors. Similar to α is the second Dirichlet parameter, β, from which the topic-specific word distributions are drawn. β values that are large tend to induce very broad topics with much overlap, whereas smaller values of β induce topics which are specific to small numbers of words. Following the empirical guidelines set forth by Griffiths and Steyvers (2004), and empirical testing performed by the author, we set the value of $\beta = 200/(\#\,\text{words})$. Given that the average corpus generally consists of ~25,000 word tokens, representing about m = 2500 unique words in about n = 1200 utterances, the value of β is generally on the order of 0.1, a value close to that used in (Rosen-Zvi et al. 2004). Tests of the AT model with β=0.1 generated results that were qualitatively similar to those with "fitted" priors, but fitted priors presented a slightly lower cross-entropy value. Thus, topics tend to be relatively word-specific. As will be shown below, values of α tend to be on the order of 1 -

5, incorporating some amount of smoothing, so the topics are not entirely author-specific. This corresponds to a number of topics between 10 and 35, depending on the specific meeting being analyzed. Tests of the model with a fixed $\alpha=5$ generated results that were also qualitatively similar to those with fitted priors, although added smoothing likely introduced some noise into analysis results, as reflected in a slightly higher cross-entropy for fixed priors for most meetings (Figure 8). The outliers are shorter meetings in which the fitted priors impose relatively more smoothing.

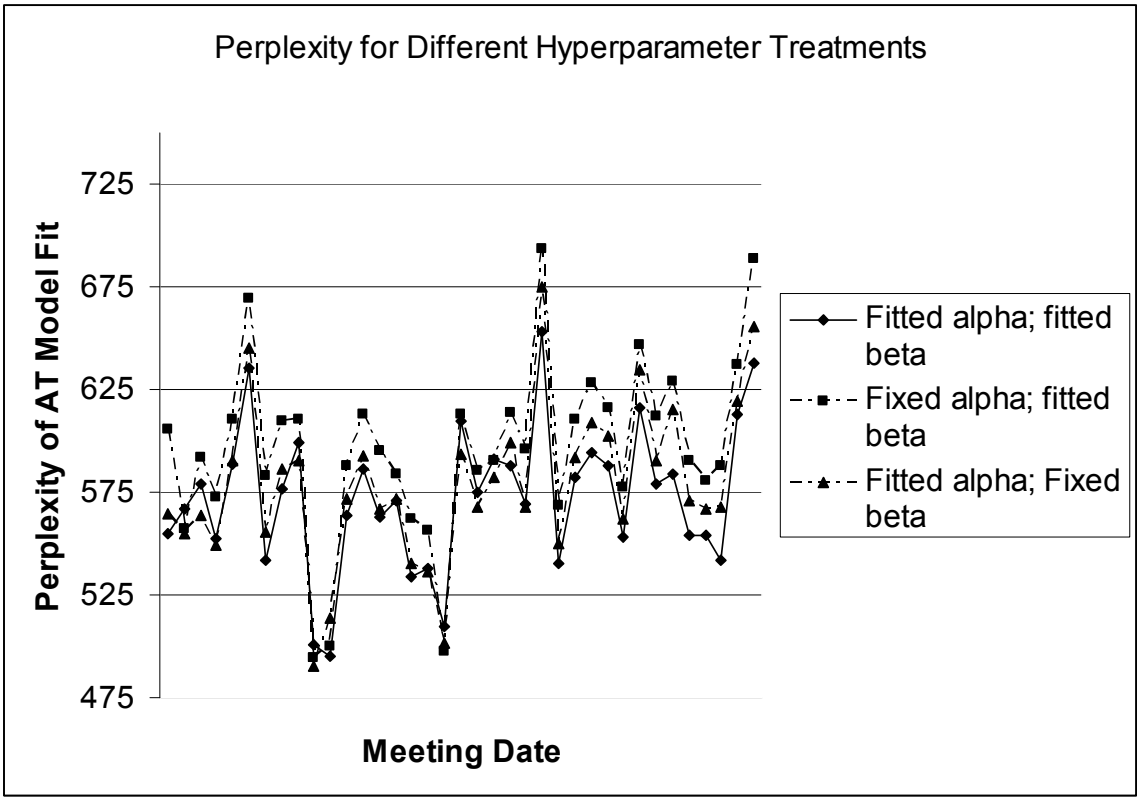


Figure 8: A comparison of perplexity values for three different hyperparameter conditions tested. Fitted priors generally have slightly lower perplexity, particularly for longer meetings.

Smaller hyperparameter values than those presented above result in even lower perplexity values. This is because the number of topics must be increased in tandem. As topics become more specific to individual words and authors (i.e hyperparameter values decrease) the number of topics required to accurately model the corpus increases, and the model has a correspondingly higher resolution. This comes at the cost of sensitivity to spurious linkages between words that might co-occur a small number of times in the corpus, without necessarily corresponding with an intuitive sense of what constitutes a topic. Ultimately, the goal of this analysis is to determine which words, and potentially ideas, speakers might have in common. If hyperparameter values are too low (i.e., topics are too author-specific or word-specific), there will be very little overlap among speakers. This means that conditional independence assumptions underlying the AT model become very strong and topics come to be defined by small numbers of relatively infrequent words (in which case many topic are required to generate a meaningful model) or entirely by speaker identity. At this point, a topic ceases to become a meaningful construct. The analysis in this thesis therefore only uses cross-entropy/perplexity sparingly as a metric of model quality to differentiate between hyperparameter schemes that have already been established in the literature. We do not try to minimize global cross-entropy/perplexity. This represents a modeling choice that departs from standard machine-learning methods – indeed it is interesting that there has been relatively little work within the topic modeling community on the appropriate choice of hyperparameters. Exceptions include hyperparameter optimization algorithms, such as those designed by (Wallach 2008) which will be discussed below. To the author's knowledge, there has been no analysis of the co-selection of topics and hyperparameters.

Committee Filtering

Our analysis primarily focuses on the voting members on an advisory panel. This decision was made because it is precisely these members whose evaluations will determine the panel recommendations. Other panel members, such as non-voting guests and consultants, are also included in the analysis because, like the voting members, they play the role of resident experts. Panel members such as the executive secretary, and consumer, patient and industry representatives are not included as part of the committee in the following analyses because they play a relatively small role in panel discussion in the meetings examined. Inclusion of these members is straightforward, and examination of their roles is left to future research.

The LSA approach demonstrates that it is often difficult to differentiate between panel members, especially since the majority of the speech during an FDA panel meeting is occupied by presentations from the sponsor and the FDA. A given voting member might speak relatively rarely. Furthermore, panel members share certain language in common including procedural words and domain-specific words that are sufficiently frequent as to prevent good topic identification. As a result, a large proportion of the words spoken by each committee member may be assigned to the same topic, preventing the AT model from identifying important differences between speakers. In a variant of a technique suggested in (Rosen-Zvi et al. 2005)[6] this problem is solved using the AT model by creating a "false author" named "committee". Prior to running the AT model's algorithm, all committee voting members' statements are labeled with two possible authors – the actual speaker and "committee". Since the AT model's MCMC algorithm randomizes over all possible authors, words that are held in common to all committee members are assigned to "committee", whereas words that are unique to each speaker are assigned to that speaker. In practice, this allows individual

[6] The author of this thesis would like to thank Dr. Mark Dredze for suggesting this approach

committee members' unique topic profiles to be identified, as demonstrated below. In the unlikely case where all committee members' language is common, half of all words will be assigned to "committee" and the other half will be assigned at random to the individual speakers in such a way as to preserve the initial distribution of that author's words over topics.

Preliminary testing of the AT Model

Preliminary tests of the AT model held the number of topics constant at 10, for each meeting analyzed. Although this is not a realistic representation of the structure of different FDA panel meetings (discussed below), these initial tests provide insight into the capabilities of the AT Model when applied to this problem domain.

Sample Output – Identifying Topics of Interest

One preliminary use of the AT model is to identify the major topics of interest for each speaker. For example, Figure 9 shows sample output of the Author-Topic model applied to the FDA Meeting held on March 4th, 2002. In particular, this is the author-specific topic distribution for one panel member. Note that a plurality of this panel member's words is confined to the topic labeled 1.

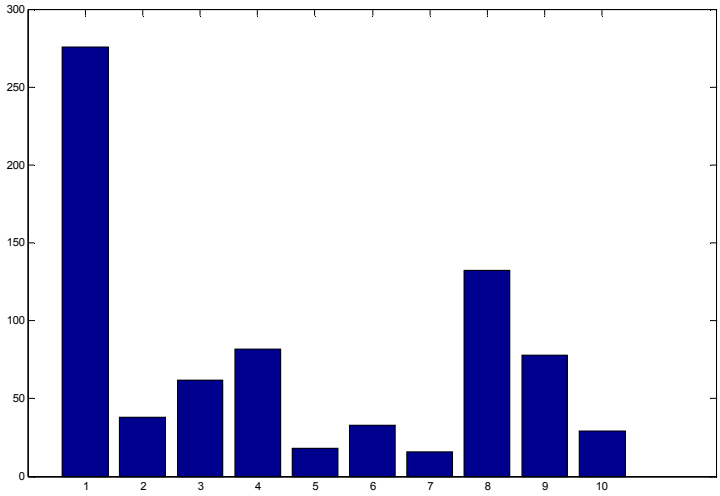


Figure 9: Sample output from the Author-Topic model run on the FDA Circulatory Systems Devices Advisory Panel Meeting for March 4th, 2002. This chart is the per-speaker topic distribution for one of the panel members.

Table 3 displays the top five most probable word stems for each topic:

Table 3: The top five word-stems for one run of the AT model on the corpus for the Circulatory Systems Devices Panel Meeting of March 4, 2002.

Topic Number	Top Five Word-Stems

1	'clinic endpoint efficaci comment base'
2	'trial insync icd studi was'
3	'was were sponsor just question'
4	'patient heart group were failur'
5	'devic panel pleas approv recommend'
6	'think would patient question don'
7	'dr condit vote data panel'
8	'effect just trial look would'
9	'lead implant complic ventricular event'
10	'patient pace lead were devic'

Within a clinical trial administered by the FDA, a device manufacturer must meet a certain set of clinical "endpoints", often defined as a proportion of a population that is free from disease or adverse events (e.g., device failure). Such trials typically have different endpoints for device safety and efficacy, both of which must be met. From this table, we can see that this panel member's major topic of interest involved questions of what was the appropriate clinical endpoint for the study in question (often debated in these panel meetings).

The use of the AT model to identify topics that are salient to each speaker can be helpful in determining how agreement is reached. Consider the meeting of the

Circulatory Systems Devices Panel held on November 20, 2003. This meeting was convened to review a PMA for approval of the Taxus ® Paclitaxel Drug-Eluting Stent, designed and marketed by Boston Scientific Corporation. Taxus was the second drug-eluting stent on the market, following the panel's decision to approve Cordis Corporation's Cypher Sirolimus-Eluting Stent one year prior. The ultimate outcome of the meeting was a consensus decision to approve the PMA. The vast majority of decisions to approve a device come with conditions of approval that the panel recommends to the FDA that the sponsor must meet before the device can be marketed. The conditions of approval for the Taxus stent were as follows:

1. The labeling should specify that patients should receive an antiplatelet regimen of aspirin and clopidogrel or ticlopidine for 6 months following receipt of the stent.

2. The labeling should state that the interaction between the TAXUS stent and stents that elute other compounds has not been studied.

3. The labeling should state the maximum permissible inflation diameter for the TAXUS Express stent.

4. The numbers in the tables in the instructions for use that report on primary effectiveness endpoints should be corrected to reflect the appropriate denominators.

5. The labeling should include the comparator term "bare metal Express stent" in the indications.

Each of these conditions may be traced to a particular voting member, or set of voting members, on the panel, using the AT model. Table 4, below, outlines the primary topics for each voting member. The top five words, identifying each

voting member's preferred topic (out of 10 total), are identified, in addition to the topic proportion – the proportion of words spoken by that voting member in the corresponding topic. Finally, each topic is assigned to a condition of approval as listed above.

Table 4: Results of the Author-Topic Model applied to a transcript of the Circulatory Systems Devices Panel Meeting of Nov. 20, 2003. Each row of this table corresponds to a different voting member. Topics correspond to conditions of approval for the final vote.

Committee Member's Medical Specialty	Major Topic of Interest (stemmed)	Topic Proportion	Corresponding Condition #
Cardiologist	'metal bare express restenosi paclitaxel'	0.36	5
Cardiologist	'physician stainless ifu steel plavix'	0.42	1
Pharmacologist	'metal bare express restenosi paclitaxel' 'materi drug interact effect potenti'	0.30 0.29	5 2
Statistician	'tabl detail denomin six number'	0.56	4
Cardiologist	'metal bare express restenosi	0.23	5

	paclitaxel'		
Cardiologist	'drug clinic present appear event'	0.23	2
Cardiologist	'angiograph reduct nine think restenosi'	0.12	<None>
Electrophysiologist	'millimet length diamet coronari lesion'	0.34	3
Surgeon	'know bit littl take present'	0.23	<None >

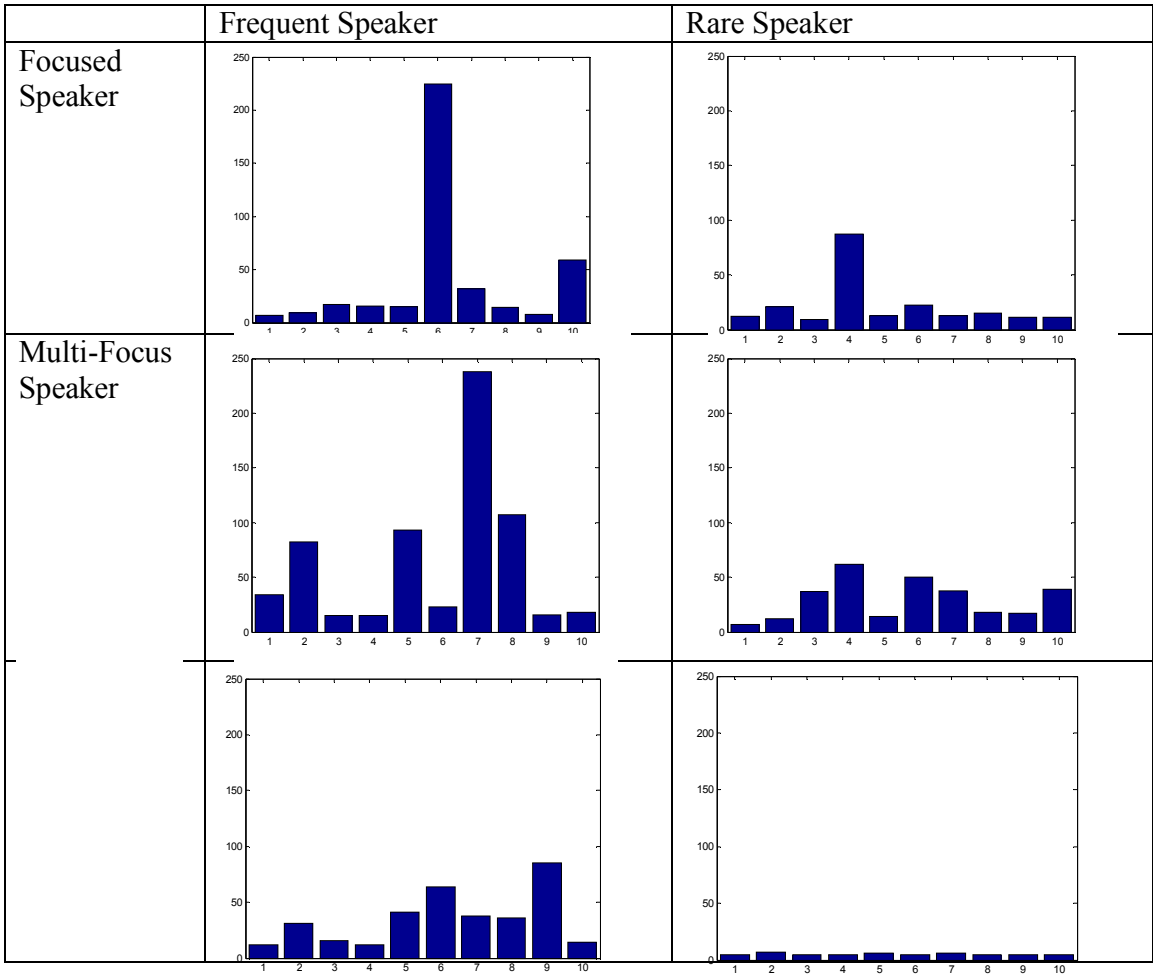
The above table shows a rough correspondence between topics of discussion and conditions of approval. This demonstrates that the AT model is able to generate author-specific topics that are meaningful to specific panel meetings. AT therefore overcomes another key limitation of LSA.

Sample Output – Identifying Roles

Aside from examining topics of interest to specific committee members, we would like to be able to examine role-based behavior. In particular, how do different panel voting members interact with one another? One preliminary insight into speaker roles may come from comparing author-specific topic distributions. Panel members who speak often and focus on one aspect of the discourse potentially display a depth of expertise and will be more likely to have their words assigned to a topic focused on that speaker. If they focus on several aspects of the discourse in concert with other speakers (e.g., if they engage in a discussion), they will tend to have their words assigned to a number of topics related to their areas of focus and potentially display a breadth of expertise. If

they do not speak often, but are focused in their area of discourse, their words will likely be assigned to topics defined by other speakers. Finally, if they speak rarely their words will be assigned uniformly at random to all topics. These different types of speakers are summarized in Table 5.

Table 5: Different types of speakers identified by the AT model. A frequent, focused speaker tends to drive topic formation, whereas a rare speaker tends to be assigned to topics defined by others. Multi-focus, or interdisciplinary, speakers may serve as mediators. These sample results have been generated from actual panel meetings.



Generation of Social Networks

We may use the output of the Author-Topic model to gain insight into the social structure of a given committee. Since the results of the Author-Topic model assign each word to a topic, we may compare topic distributions across speakers. In particular, if two speakers' words are assigned to the same topic frequently, we say that they are "linked". The definition of a link is another modeling choice. Early versions of this algorithm considered two authors to be linked if, in a model with ten topics, they had at least one topic in which they both spoke more than 20% of the time. In the current version, speakers are linked together if they

commonly use the same topics of discourse. In particular, we construct an Author-Topic matrix, **A**, with entries equal to 1 where that author uses that topic, and entries equal to 0 otherwise. This matrix, when multiplied by its transpose (**A * A'**) yields a linkage pattern among speakers. This may be interpreted as a social network (Wasserman and Faust 1994). A more rigorous definition of linkage between speaker-pairs is to be found below. Using authors as nodes, and the links derived from their topic distributions as edges, we may generate an author-topic graph. Because we are only interested in the voting members (and committee chair) we analyze only the subgraphs consisting of the nodes and edges linking these members, such as that shown in Figure 10.

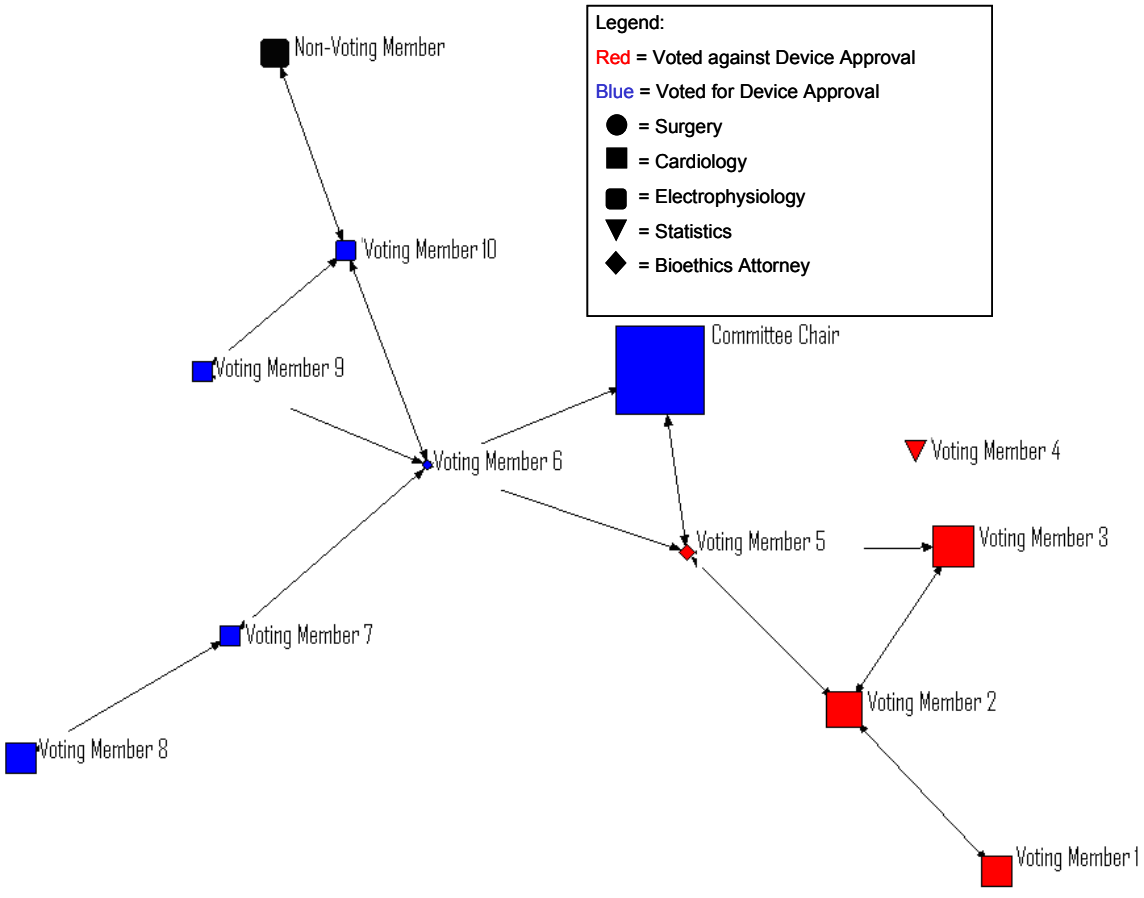


Figure 10: A graph of the meeting of the FDA Circulatory Systems Devices

Advisory Panel Meeting held on March 5, 2002. Node size is proportional to the number of words spoken by the corresponding speaker. Random seed = 613. Graphs were generated using UCINET.

Different schemes for determining topic membership yield different networks. For example, the binomial statistical test might be seen as a more principled way of determining topic membership. The binomial statistical test operates by examining the cumulative distribution of the binomial probability mass function, given by

$$\Pr(K = k) = \binom{n}{k} p^k (1-p)^{n-k} \quad \textbf{(8)}$$

Under this scheme, an author is assigned to a topic if the cumulative probability that that author used k out of n words in a given topic is less than 0.05/b, where b is the Bonferroni significance level correction factor. Given a authors, b = a * (a-1) / 2, since one comparison is being made for each pair of authors. Unlike the uniform 20% cutoff used above, the binomial cutoff accounts for the total number of words a given speaker contributes to a given topic. Therefore panel members who speak rarely are less likely to be linked. A sample social network from this scheme is shown in Figure 11, for the same meeting as in Figure 10.

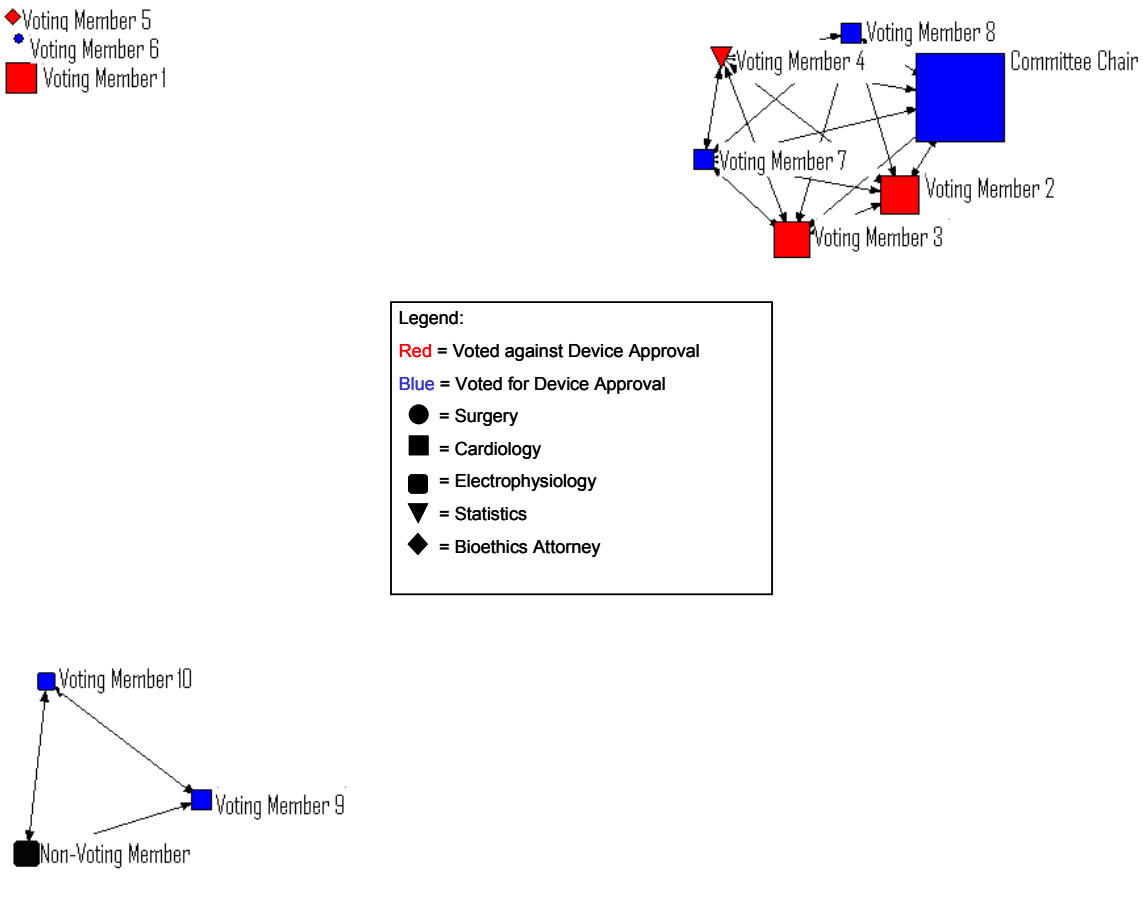


Figure 11: Social network of the FDA Circulatory Systems Devices Advisory Panel Meeting held on March 5, 2002. Threshold value is determined using the binomial test described above. Node size is proportional to the number of words spoken by the corresponding speaker. Random seed = 201.657. 2100th draw from MCMC algorithm. Graphs were generated using UCINET. This iteration shows the presence of two separate discussion groups. Note that voting members 5 and 6, both bridging members in Figure 10, are now disconnected. This is due to their small number of words contributed.

As above, each social network generated using this scheme is the result of one MCMC iteration. Multiple iterations, when taken together, form a probability distribution over a set of possible Author-Topic assignments, and therefore, connectivity patterns. We can expect that different iterations of the MCMC algorithm will yield drastically different graphs. For example, the results of a second draw from the same MCMC chain that yielded Figure 11 is shown in Figure 12.

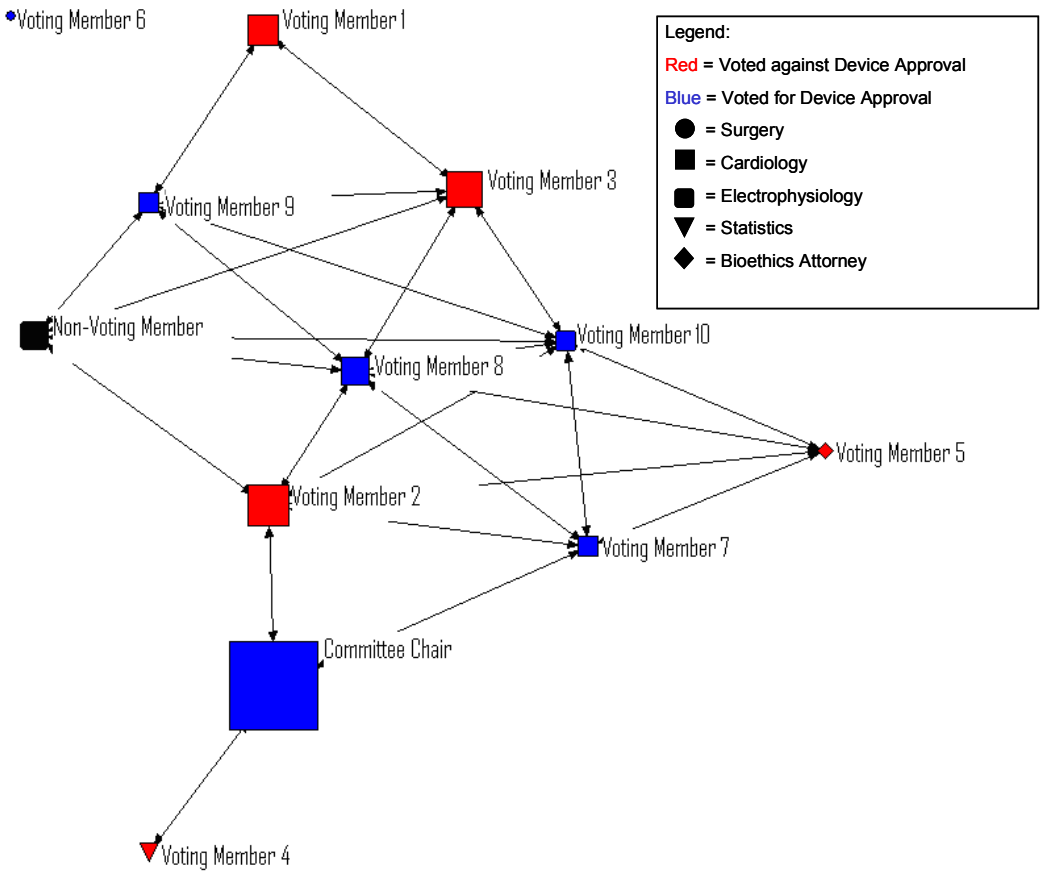


Figure 12: A second iteration of the meeting of the FDA Circulatory Systems Devices Advisory Panel

Meeting held on March 5, 2002. Threshold value is determined using the binomial test described above. Node size is proportional to the number of words spoken by the corresponding speaker. Random seed = 201.657. 2200th draw from MCMC algorithm. Graphs were generated using UCINET

The high variability among draws from the MCMC algorithm again suggests that links should be differentially weighted – some links appear in virtually all iterations, whereas other links appear in relatively few iterations. Averaging over multiple MCMC iterations enables a social network to be created with weighted links, where the weight of each link is proportional to its frequency of occurrence among iterations. Examples of this may be found in Figure 13 and Figure 14, corresponding to constant and binomial threshold conditions, respectively.

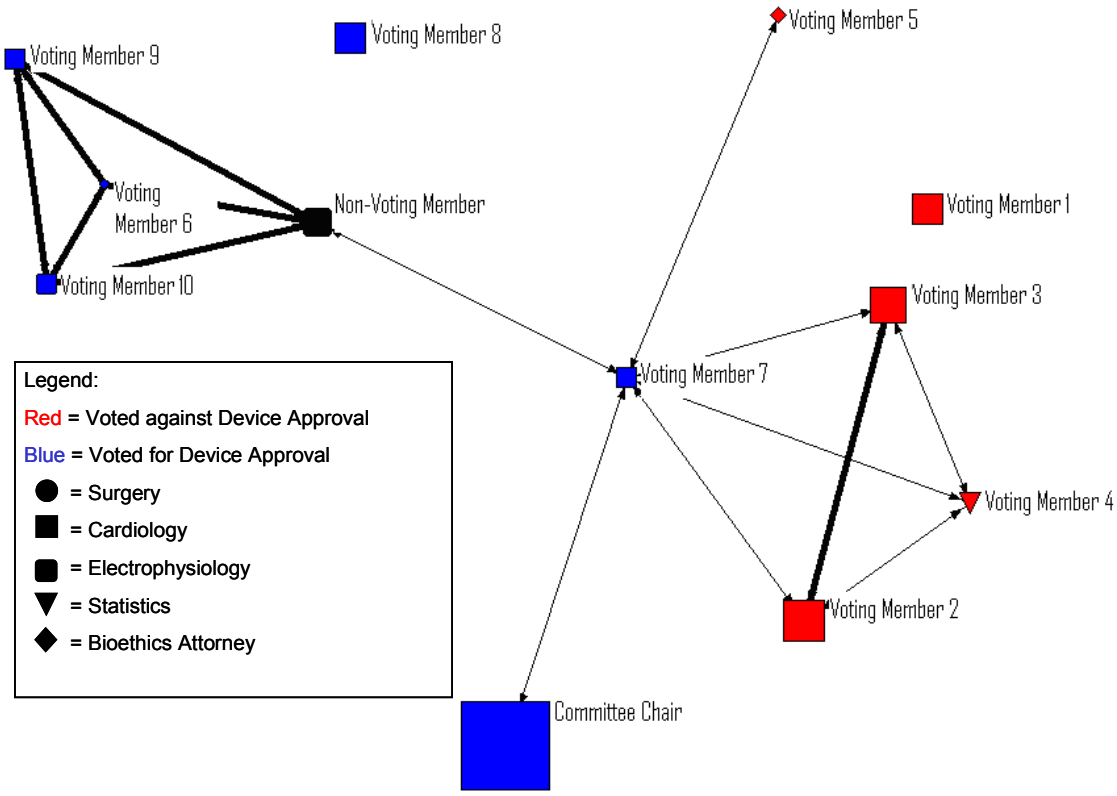


Figure 13: Average of 200 graphs for the meeting of the FDA Circulatory Systems Devices Advisory Panel Meeting held on March 5, 2002. A heavy line indicates a strong link (linked in >100 graphs). A light line indicates that the speakers are linked more frequently than the global average of all speakers. Spurious links have been eliminated.

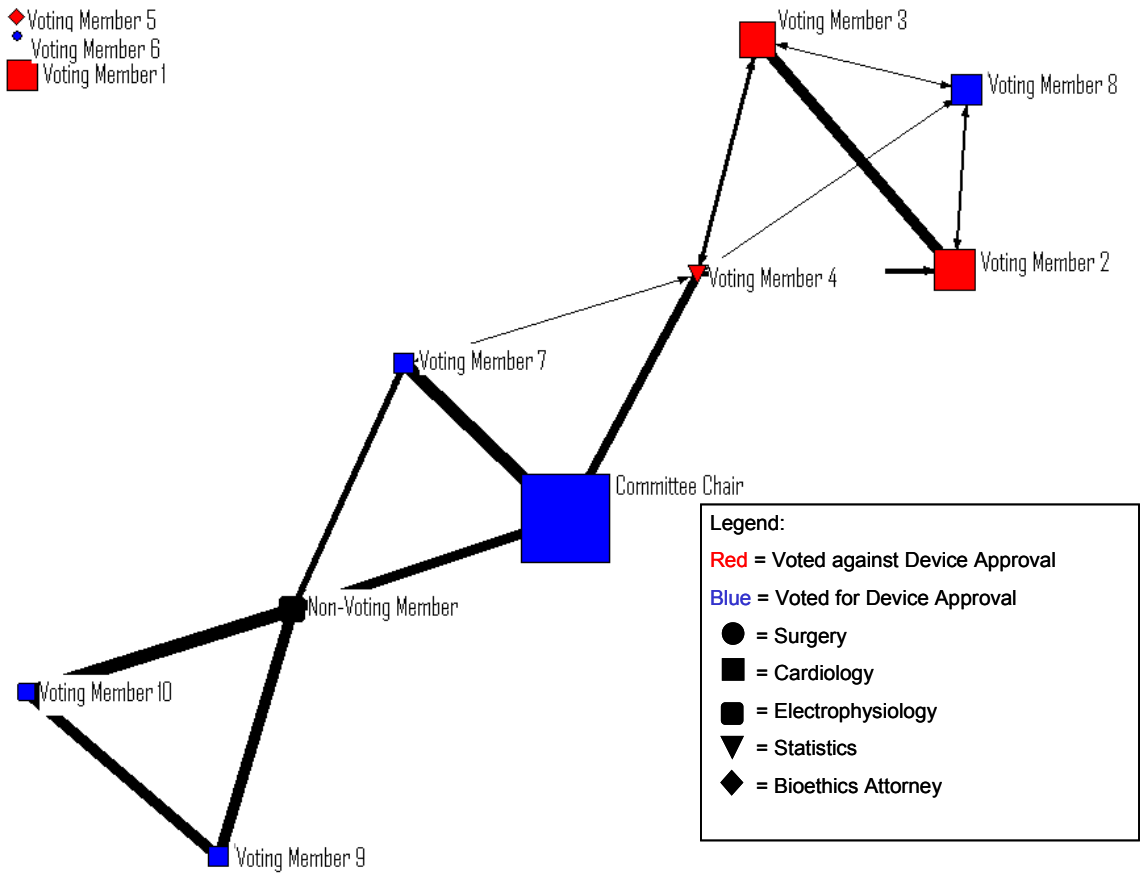


Figure 14: Average of 200 iterations for the meeting of the FDA Circulatory Systems Devices Advisory Panel Meeting held on March 5, 2002. Iterations use a binomial threshold value for each of ten topics. Heavier lines indicate stronger links (linked in >100 iterations), whereas lighter lines indicate weaker links (> than the global average). All links shown are stronger than the global average of all speakers. Remaining links have been deleted.

Despite differences in the locations of rare speakers, Figure 13 and Figure 14 have similar overall structures. For example, both figures display a structure that tends to group together those speakers who voted similarly. This is encouraging for testing hypotheses about voters who speak the same way tend to vote the same way.

The difference in the two figures highlights the differences between the two threshold conditions. The constant threshold condition tends to favor speakers who speak infrequently, such as voting members 5 and 6. Because of their relatively small numbers of words, it is harder for these speakers to achieve statistical significance using the binomial test, and so they are less likely to be linked. On the other hand, the constant threshold condition requires more words to establish a link to a frequent speaker, compared to a binomial threshold.

Comparison of multiple cases

The case in the previous section demonstrated a preliminary method for how social networks can be built. Later in this section, we will discuss how to refine this method. Nevertheless, it is instructive to perform some preliminary analyses of the capabilities of these early networks.

Grouping by medical specialty?

Network representations of some meetings display voting along institutional lines more clearly than do others. For example, Figure 15 and Figure 16 show a strong grouping by medical specialty. In particular, surgeons and internal medicine experts (cardiologist and pharmacologists) seem to form two different parts of the same graph.

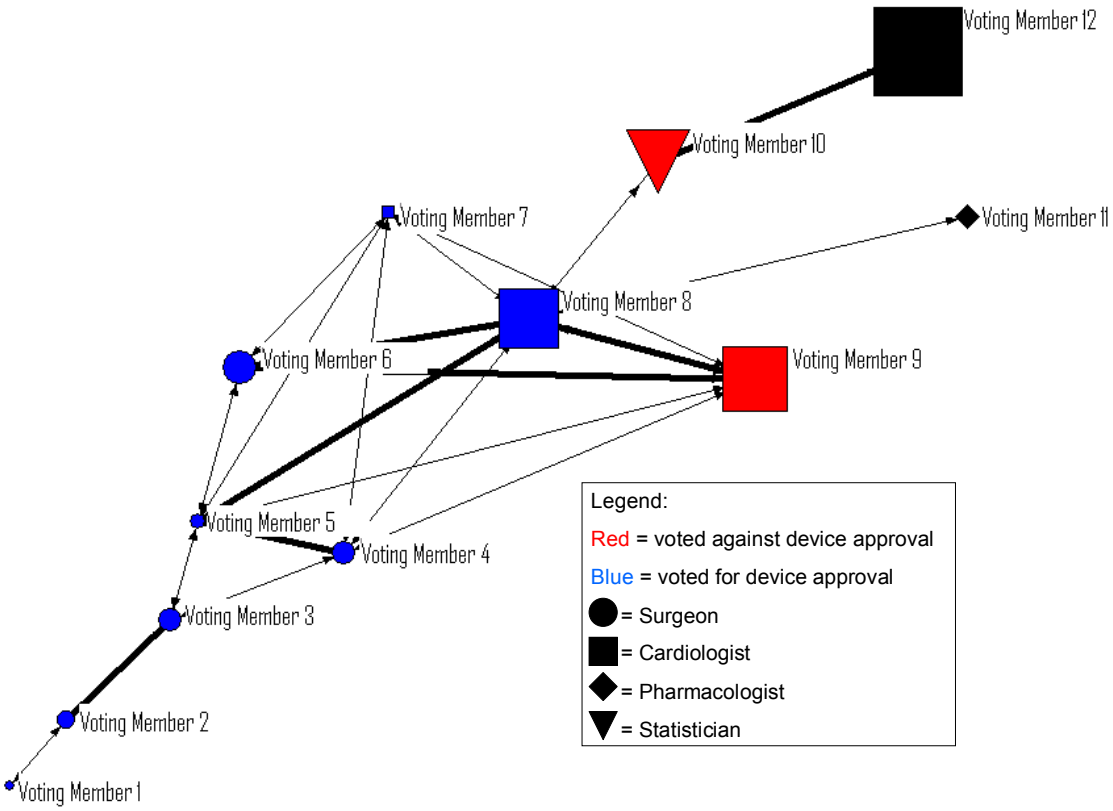


Figure 15: Average of 200 iterations for the meeting of the FDA Circulatory Systems Devices Advisory Panel Meeting held on January 13, 2005. Iterations use a constant threshold value for each of ten topics. A heavy line indicates a strong link (linked in >100 iterations). A light line indicates that the speakers are linked more than the global average of all speakers. Remaining links have been deleted.

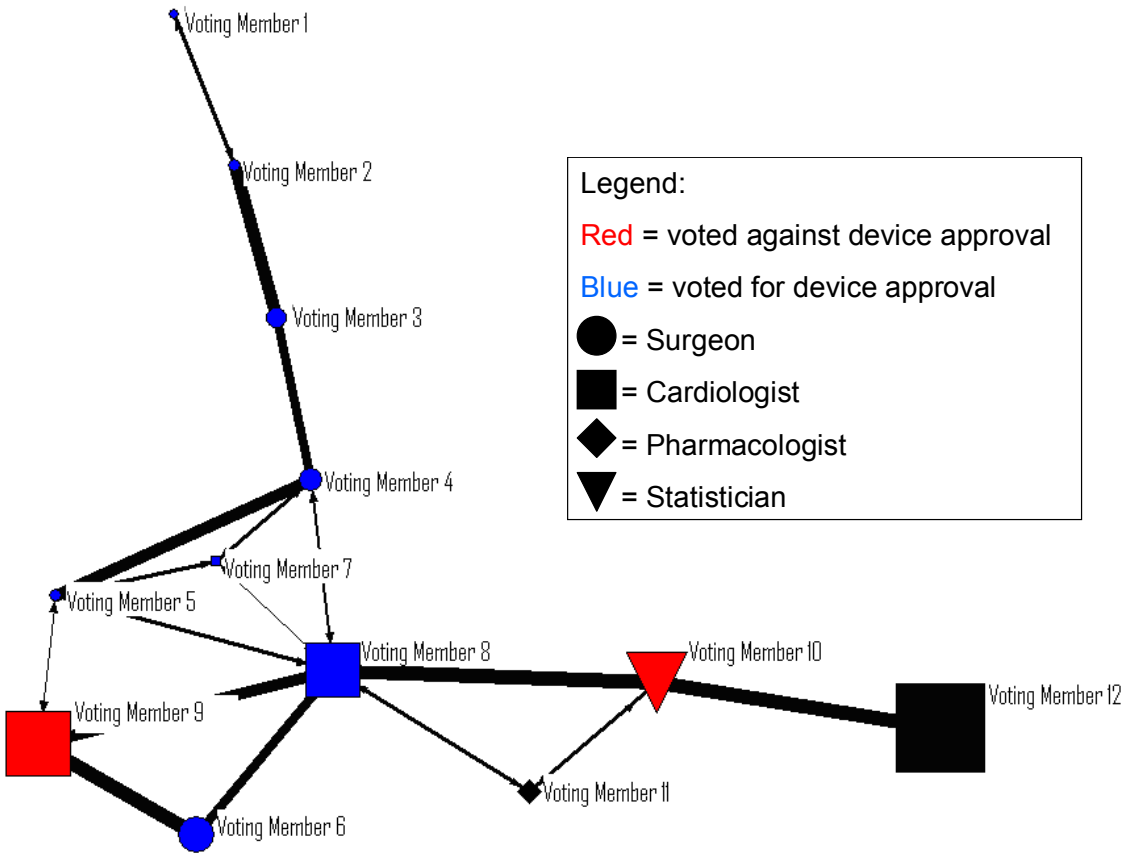


Figure 16: Average of 200 iterations for the meeting of the FDA Circulatory Systems Devices Advisory Panel Meeting held on January 13, 2005. Iterations use a binomial threshold value for each of ten topics. Heavier lines indicate stronger links, whereas lighter lines indicate weaker links. All links shown are stronger than the global average of all speakers. Remaining links have been deleted.

Both of these figures place Voting Member 8 in the most central position on the graph of committee voting members. Both graphs also show a potentially

important role for Voting Member 6 who is graphically closer to the cardiologists while still voting with the other surgeons. It may be significant that Voting Member 6 was Canadian whereas all other surgeons were employed at hospitals in the United States. Furthermore, Figure 16 recognizes strong links between Voting Members 9 and 10 to Voting Member 8. This is consistent with a reading of the meeting transcript that indicates that Voting Member 8 shared many of the concerns of her colleagues, despite ultimately voting with the surgeon majority. Voting Member 12, who abstained from voting, is strongly linked to Voting Member 10, consistent with his background as a clinical trial designer who would be interested in both the clinical and the statistical elements of the analysis. It is interesting to note that both figures also display long "tails" of surgeons, who seem to have voted as a bloc.

The above results indicate that, at least in some cases, medical specialty might have some predictive value for voting outcomes. Further analysis in Chapter 5 is aimed at attempting to confirm or deny this hypothesis. Of particular interest are those panel members who are linked across specialty boundaries. These individuals might possess a skill set or personal inclination that enables them to mediate between or learn from panel members in other specialties. This might be associated with a breadth of expertise.

Selection of Number of Topics

Without any knowledge of the content of a particular meeting corpus, it is difficult to choose an appropriate number of topics, T, *a priori*. Given hyperparameter values, as defined above, we may use perplexity as a metric for choosing T. Ideally, T would be chosen so as to be as small as possible (i.e., maximum dimensionality reduction) while still constituting a good model fit.

The number of topics is chosen independently for each transcript as follows: 35 AT models are fit to the transcript for t = 1 … 35 topics – given the

hyperparameter values as defined above, we found that 35 topics was an appropriate upper bound since, as the number of topics increases, model cross-entropy becomes asymptotically smaller. When fixed values of α are used, there is a unique minimum in the function relating perplexity to number of topics. Griffiths and Steyvers (2004) report a similar unique minimum for fitted values of α with fixed values of β, although they tested topics in increments of 100 – their analysis did not focus on finding a model that fit a comparatively precise number of topics within the neighborhood of the minimum value. In principle, given a sufficiently large number of topics, the perplexity would begin to increase at a relatively mild slope as the model starts over-fitting. Lacking such a unique minimum here, we choose the minimum number of topics such that the cross-entropy values are statistically indistinguishable from larger numbers of topics. Thus, for each model, 20 independent samples are generated from one randomly initialized Markov chain after a burn-in of 1000 iterations. Sample independence is guaranteed by introducing a lag of 50 iterations between each sample (lags as large as 100 iterations were tested, yielding qualitatively similar results). We find the smallest value, t_0, such that the 95th percentile of all samples for all larger values of t is greater than the 5th percentile of t_0. Given fitted priors of the sort recommended by Griffiths and Steyvers (2004), the asymptotic behavior displayed in Figure 17 is typical of AT Model fits. We set the value of $T = t_0 + 1$ so as to ensure that the model chosen is well beyond the knee in the curve, and therefore in the neighborhood of the minimum perplexity.

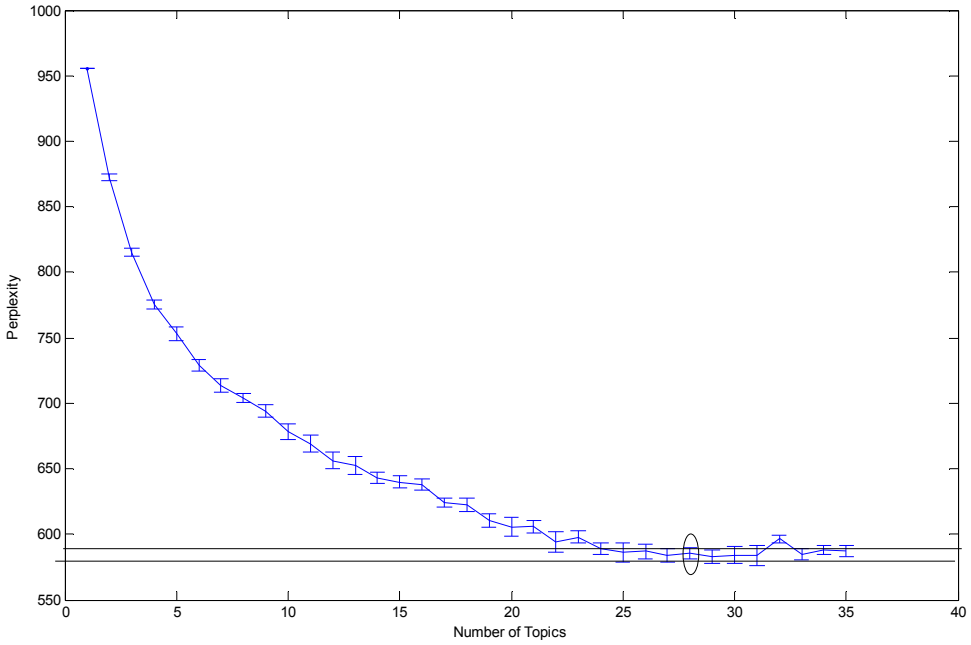


Figure 17: Perplexity vs. number of topics for the meeting of the FDA Circulatory Systems Devices Panel held on July 9, 2001. T, the number of topics, is equal to 28, using the procedure described above. Horizontal lines indicate the 5th and 95th percentiles for perplexity for a 27 topic model fit.

Once the number of topics has been chosen, a T-topic AT Model is again fit to the transcript. Ten samples are taken from 20 randomly initialized Markov chains, such that there are 200 samples in total. These form the basis for all subsequent analysis.

Future work in parameter selection could focus on incorporating these parameters into a fully Bayesian framework. For example, (Wallach 2008)

presents a hyperparameter optimization algorithm[7] that, when used with the AT model on FDA panel meeting transcripts, generates non-symmetric hyperparameter values of α that are roughly two orders of magnitude smaller than those used in this analysis. Values of β remain roughly similar to those presented here. In order to minimize perplexity for these hyperparameter values, the number of topics must be increased by roughly one order of magnitude to about 300 per meeting. These topics are extremely specific, such that there is virtually no overlap between authors. This one extreme implementation of the AT model takes the modeling assumptions to its limits and, although perplexity is absolutely minimized, the modeling assumptions so dominate the analysis as to render it useless for the applications intended in this thesis – namely comparison of topic overlap between speakers.

Selection of Network Cutoff

Network Construction

We would like to develop a principled way to determine what constitutes a link within a given model iteration. As noted above, we would like to link together speakers who commonly use the same topics of discourse. In particular, we examine each author-pair's joint probability of speaking about the same topic.

$$P(X_1 \cap X_2) = \sum_{i}^{T} P(Z = z_i \mid X_1) * P(Z = z_i \mid X_2) \qquad \textbf{(9)}$$

We would like to be able to construct an Author-Author matrix, $\boldsymbol{\Delta}$, with entries equal to 1 for each linked author pair, and entries equal to 0 otherwise. Note that this is different from the author-topic matrix, **A**, noted above.

[7] Wallach's algorithm requires a prior over hyperparameter values. The results of this test used an "improper prior" – i.e., a prior set equal to zero. This is equivalent to a fully data-driven hyperparameter selection process that, empirically, has a tendency to over-emphasize the independence of authors for the purposes of the analyses performed in this work.

Author-Author Matrix Determination

The AT model outputs an Author-Topic matrix, **A**, that gives the total number of words assigned to each topic for each author. This information must be reduced to the **Δ** matrix identified above. The form of the author-topic model makes an explicit assumption regarding an author's prior distribution over topics. This value is expressed by the hyperparameter α. Given the number of topics fit to a particular model, we may use the value of α to generate a set of *a priori* author-specific topic distributions. These, in turn, can be input into the equation above in order to generate a prior distribution for any given author-pair's link probability. Such a distribution is shown in Figure 18

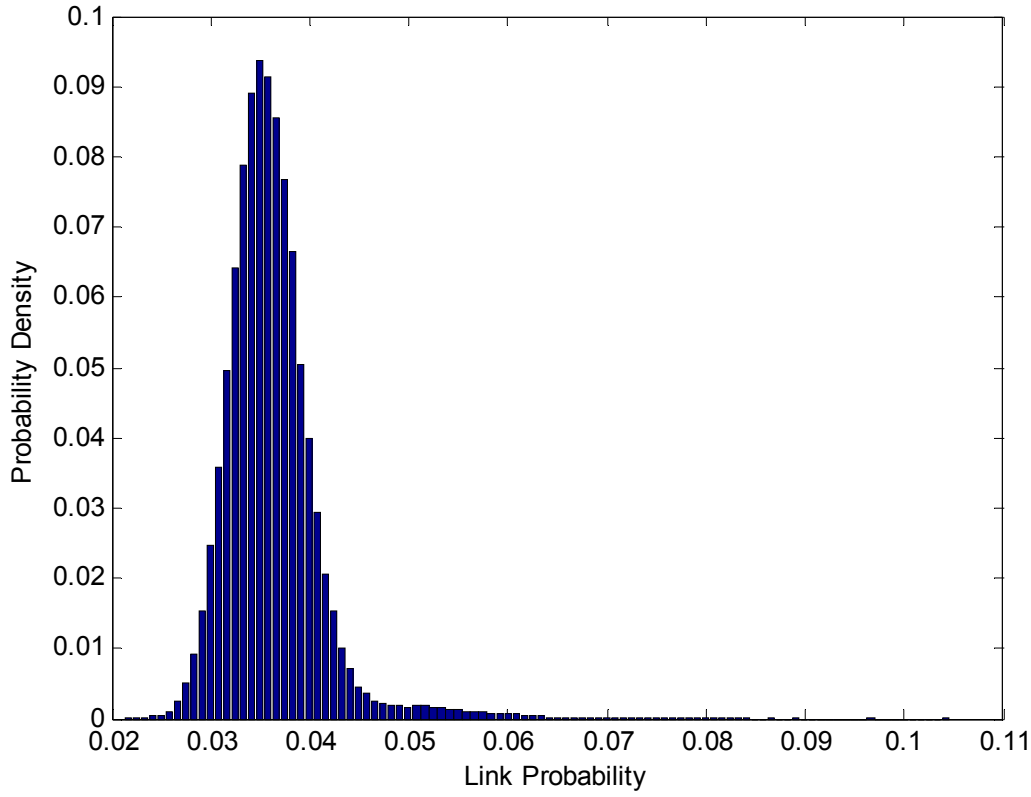


Figure 18: *A priori* probability distribution for links between speakers in the April 21, 2004 meeting with 28

topics. The median of this distribution is 0.0356; whereas 1/28 = 0.0357. The assumption of a symmetric Dirchlet prior distribution implies that this distribution holds for all speakers until it is updated with data observed from the transcripts.

In practice, the median value of this distribution becomes arbitrarily close to 1/(# topics). Therefore, within one iteration we assign a link if the observed probability that a given author pair discusses the same topic is linked exceeds 1/(# topics). In other words, it is more likely than not that the author-pair is linked. If there are 10 topics, we would expect every author-pair to have a 10% probability of being linked, *a priori*. We consider an author pair to be linked within a given model iteration if that pair's joint probability exceeds what we would expect under a uniform distribution. This scheme allows network construction to adapt to changing numbers of topics.

As before, we average over multiple MCMC iterations to enable a social network to be created with weighted links, where the weight of each link is proportional to its frequency of occurrence among iterations. Nevertheless, the variability among draws from the MCMC algorithm suggests that links should not be weighted. Histograms of the distribution of these link frequency values tend to show a bimodal structure (see Figure 19) suggesting that a description of author pairs as either connected or not connected is appropriate.

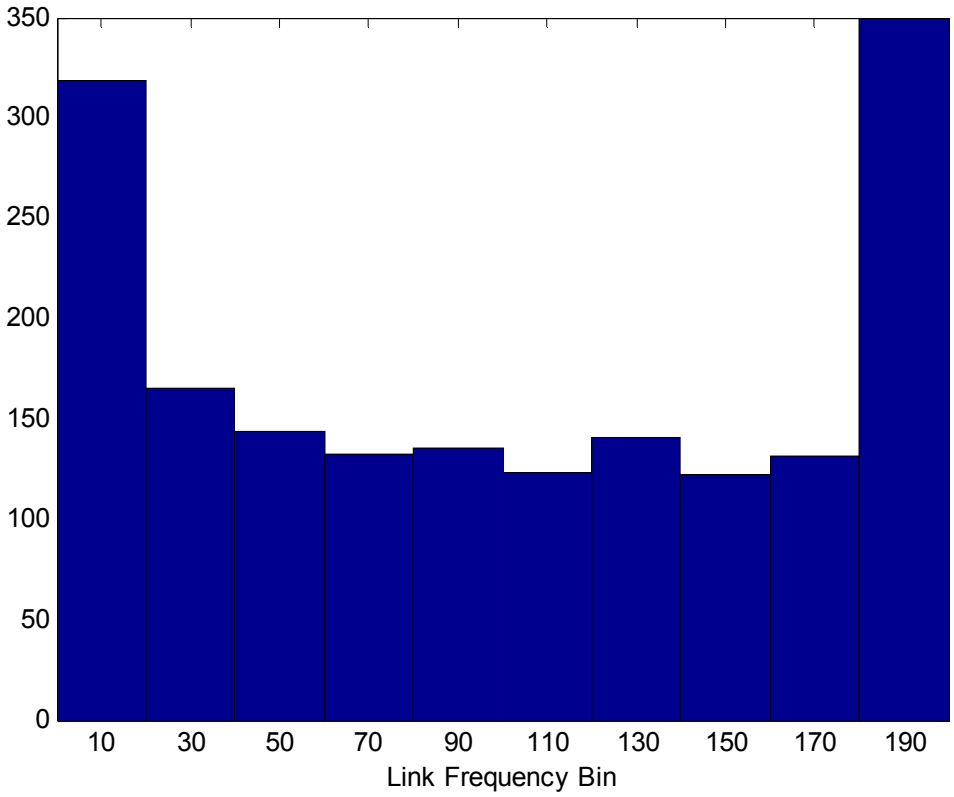


Figure 19: Sample histogram of linkage frequency for an FDA Advisory Panel meeting of April 21, 2004. The horizontal axis is the link weight (i.e., the frequency with which author-pairs are connected over 200 samples from the AT model). The vertical axis is the link frequency of links with the weight specified by the abcissa (i.e., the number of author-pairs that are connected with the frequency specified by the abcissa). Note the existence of two modes located at the extremes of the distribution.

The final challenge in constructing a network is determining where to establish the cutoff beyond which we accept that a pair of speakers is linked.

Bonferroni Cutoff Criterion

Two authors are considered to be linked in a network if they are more likely to be connected by an edge in a given sample iteration than not. Since there are 200 samples from which a link might be inferred, we would like to establish a cutoff value that is consistent across networks. The largest committee in our sample of 37 FDA advisory panel meetings possesses 15 potential voting members (not including the committee chair). Therefore, the largest network has 15*14/2 = 105 potential links among voting members. Each potential link must be tested in order to determine if it occurs more frequently than would be expected by chance. Lacking any prior information on link probabilities, we assume that a given speaker has no predisposition towards either linking or not linking. Therefore, we would expect that a randomly chosen pair of speakers would be linked 100 times out of 200. We would like to know if a given pair's link frequency is higher than what we would expect under a uniform distribution across conditions of linkage and no linkage. The binomial test may be used for precisely this sort of analysis. Furthermore, given that we are testing up to 105 different independent potential links, the p-value for this test should be subject to a Bonferroni correction. Using a binomial test, and a family-wise error rate of p=0.05, a given author pair must be linked at least 125 times out of 200 samples to be considered more frequently linked than we would expect by chance. This is the criterion that we use for all of the results presented in Chapter 5.

Sample Networks

The results of the analysis below anecdotally support the assertion that language and medical specialty are correlated. Nevertheless, some meetings display voting along institutional lines more clearly than do others. For example, Figure 20 and

Figure 21 show a strong grouping by medical specialty. Such clustering relationships are reminiscent of work in social psychology, in particular Dynamic Social Impact Theory (Nowak, Szamrej et al. 1990; Latane 1996).

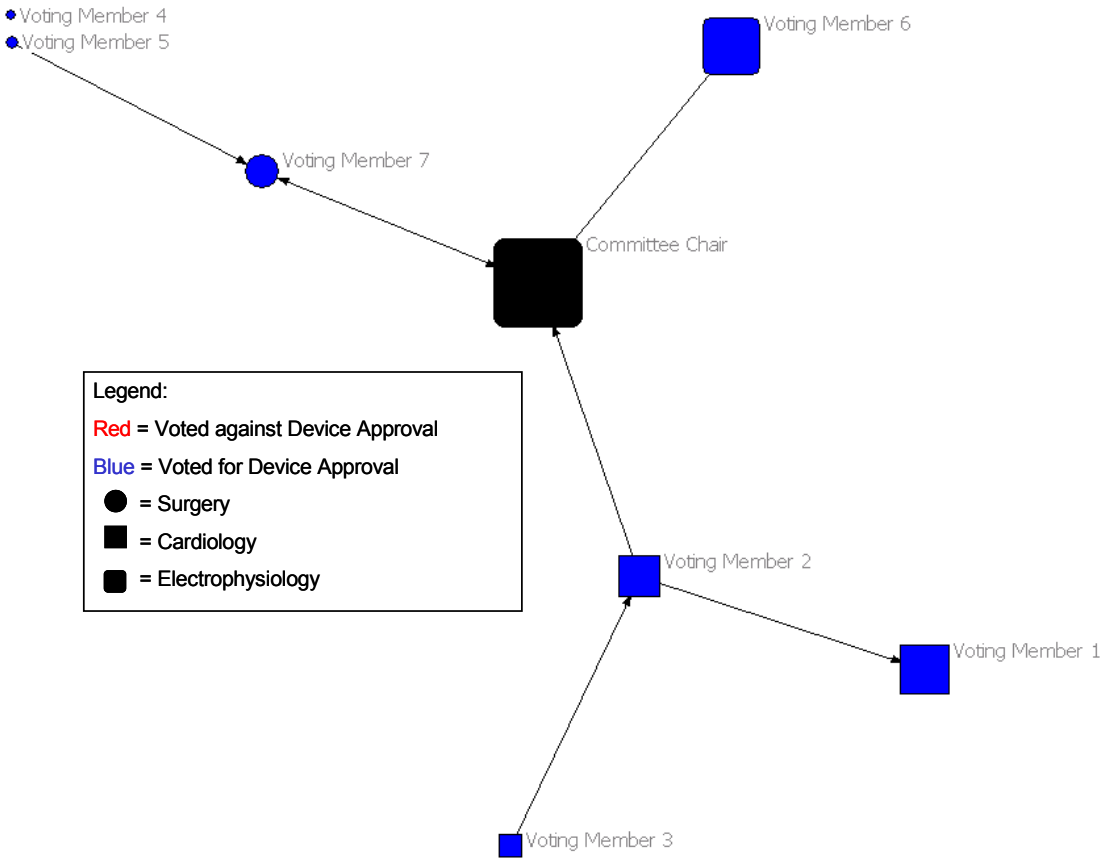


Figure 20: Graph of the FDA Circulatory Systems Advisory Panel meeting held on December 5, 2000. This meeting yielded a consensus approval of the medical device under analysis. Node shape represents medical specialty. The committee chair is in black.

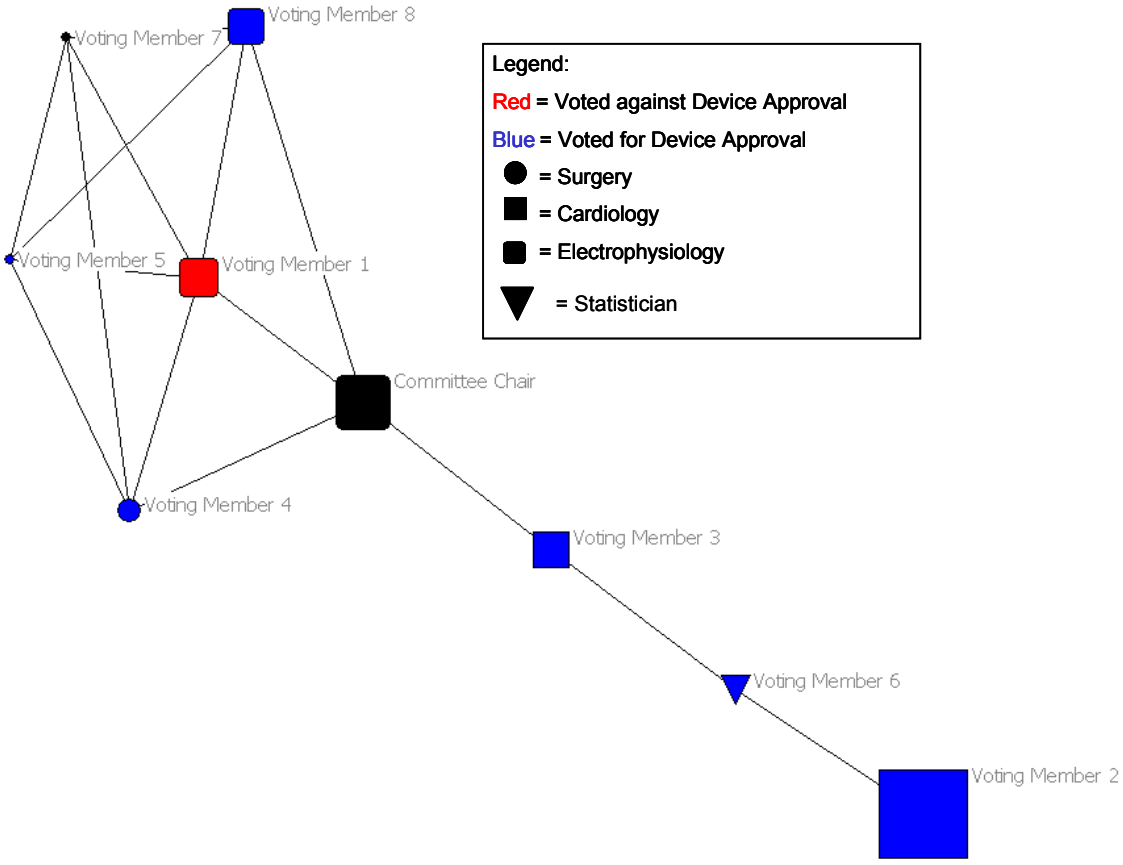


Figure 21: Graph of the FDA Circulatory Systems Advisory Panel meeting held on October 27, 1998. This meeting yielded an approval of the medical device under analysis, with only one dissenter (in red). Node shape represents medical specialty. The committee chair is labeled and did not vote. The voter in black was not present for the vote.

Grouping by Votes

In situations where the panel's vote is split, the method described in this paper can often isolate voting cliques (see Figure 22 and Figure 23). In some meetings, medical specialty and vote are aligned. This is the case in Figure 22. In this

meeting, all surgeons voted against device approval, whereas most cardiologists voted in favor. Radiologists' votes were split evenly between the two. In others, such as Figure 23, there is a weaker correspondence. Both graphs show members of the same voting coalition to be connected. It is interesting that the device analyzed in the meeting represented by Figure 23 would not have been used by the vast majority of the medical specialties represented on the panel. That panel members interacted more frequently across boundaries on this device suggests a context-dependence for specialty grouping. Furthermore, both of these meetings were quite long, (approximately 10 hours) suggesting that panel members may have taken more time to learn from one another in the face of uncertain data.

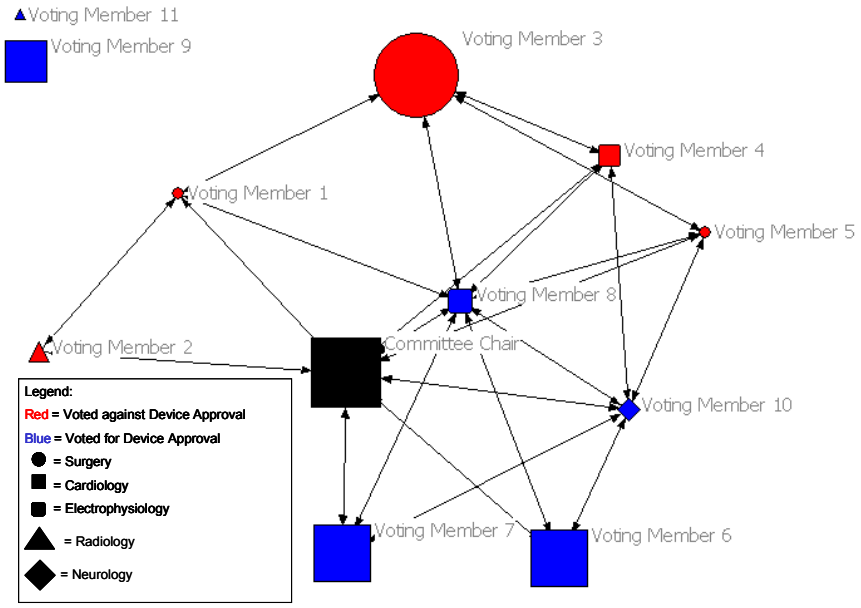


Figure 22. Graph of the FDA Circulatory Systems Advisory Panel meeting held on April 21, 2004. This meeting yielded an approval of the medical device under analysis, although the panel was split (blue, in favor; red against). Node shape

represents medical specialty. The committee chair is in black.

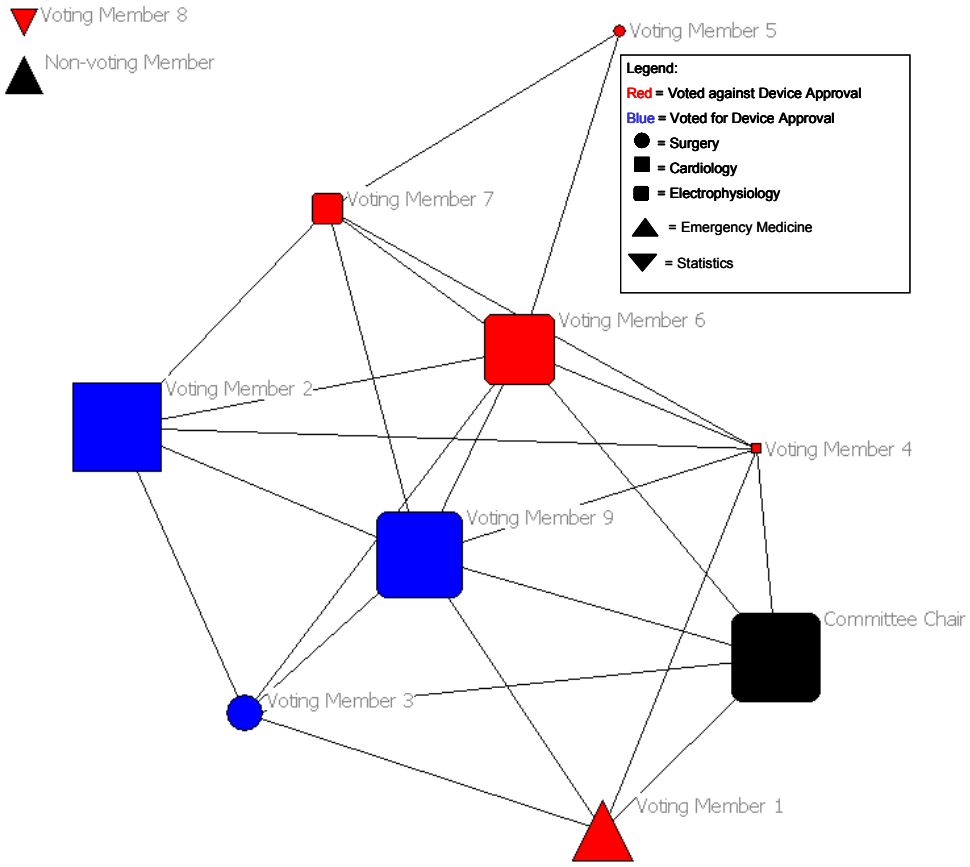


Figure 23: Graph of the FDA Circulatory Systems Advisory Panel meeting held on June 6, 1998. This device was not approved. Node shape represents medical specialty. The committee chair is in black. Non-approval votes are in red; approval votes are in blue. In this meeting, vote is not correlated with medical specialty.

In many of the cases for which graphs were generated, connectivity patterns could be understood using vote or specialty information alone. Chapter 5 will present an analysis exploring the relation between network connectivity and cohesion by vote or specialty.

Other Factors

On June 23, 2005 the Circulatory Systems Devices Panel held a meeting to determine whether a particular device should be approved for a Humanitarian Device Exemption. Such a meeting almost surely appeals to a sense of personal ethical responsibility that transcends medical specialty. In situations such as these, we might expect that individual votes and connectivity patterns will be more idiosyncratic and exhibit less coherence. Figure 24 shows the connectivity pattern for this meeting. Note that this graph cannot be as easily partitioned by vote or by medical specialty confirming the idea that the evaluation is independent of medical specialty and suggesting that voting blocs are not operative in this special case. That voting blocs are not connected in this graph suggests that dialogue may not be effective in changing preferences in the face of ethical or value-based decision-making.

.

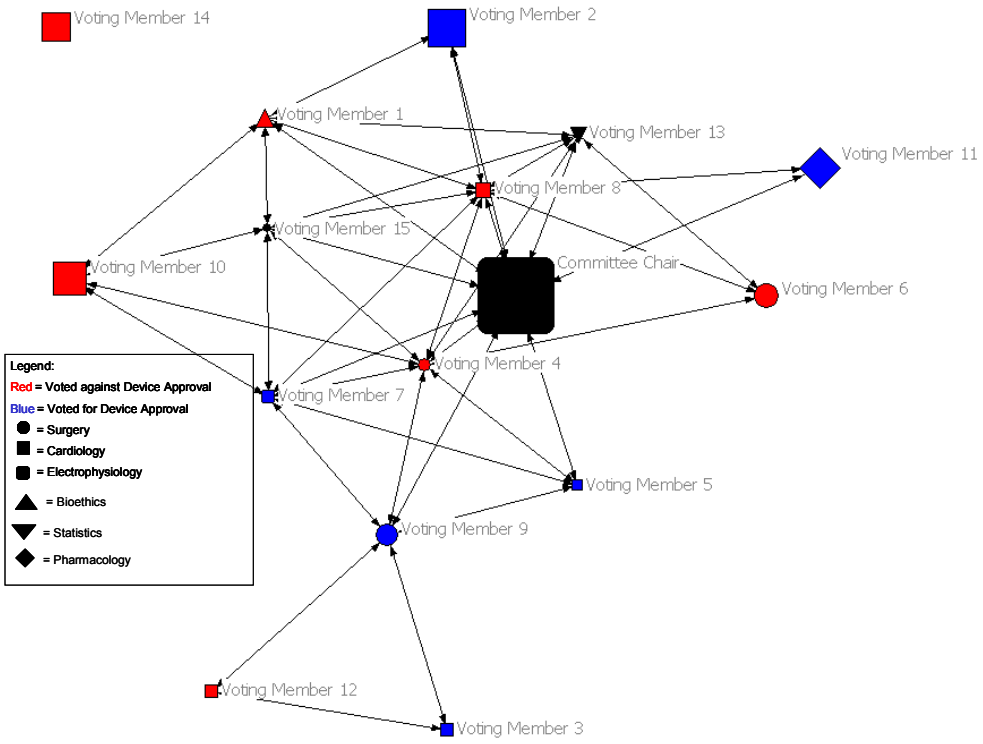


Figure 24: Graph of the FDA Circulatory Systems Advisory Panel meeting held on June 23, 2005. Node color represents the vote (red is against humanitarian device exemption, blue is in favor of humanitarian device exemption, black is abstention. The committee chair is also black. Node shape represents medical specialty.

Explicitly Representing Uncertainty in Graphs

Future work can focus on explicit representation of link uncertainty between speakers. For example, given the prior background link distribution shown in Figure 18, we might ask how much more likely is a particular link to occur given its author-pair-specific distribution. An example of a prior link distribution

compared to its posterior is shown in Figure 25, for a very strongly linked pair of authors.

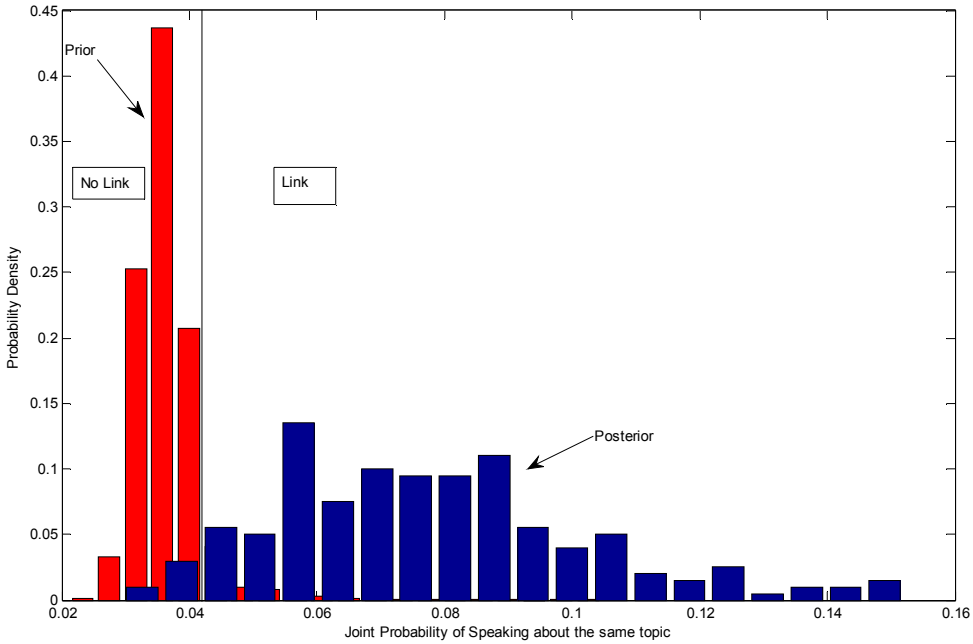


Figure 25: Comparison of prior and posterior distribution of link probabilities for two strongly-linked voting members during the April 21, 2004 meeting. An ideal observer would place the link probability threshold around 0.04, indicating that a joint probability greater than this value would signal a link with very high likelihood.

Given a pair of probability distributions we would like to determine how likely it is that a random sample is drawn from the posterior distribution, as opposed to the prior distribution. If the posterior distribution is treated as a signal and the prior distribution is treated as noise, we may formulate this problem as one of

signal detection. Solving this problem first requires setting a threshold, Θ. An *ideal observer*, i.e., one who assigns equal weight to false positives as to false negatives, would set Θ at the point of intersection between the curves representing the posterior and prior distributions. All observations that are greater than Θ would be considered evidence of a link. Given Θ (which might be calculated using the distributions described above), we can calculate the *likelihood-ratio* or *signal-to-noise ratio* of a given link. This is simply the probability of a correct detection divided by the probability of a false positive:

$$LR_{1,2} = \frac{\int_{\Theta}^{+\infty} P(x_1 \cap x_2 \mid w, \alpha, \beta)}{\int_{\Theta}^{+\infty} P(x_1 \cap x_2 \mid \alpha, \beta)} \quad \textbf{(10)}$$

$LR_{1,2}$ may serve as an edge weight on a graph between nodes representing speakers 1 and 2. For values of $LR_{1,2} > 1$, a link is more likely than not. Examples of such graphs are shown below in Figure 26 and Figure 27.

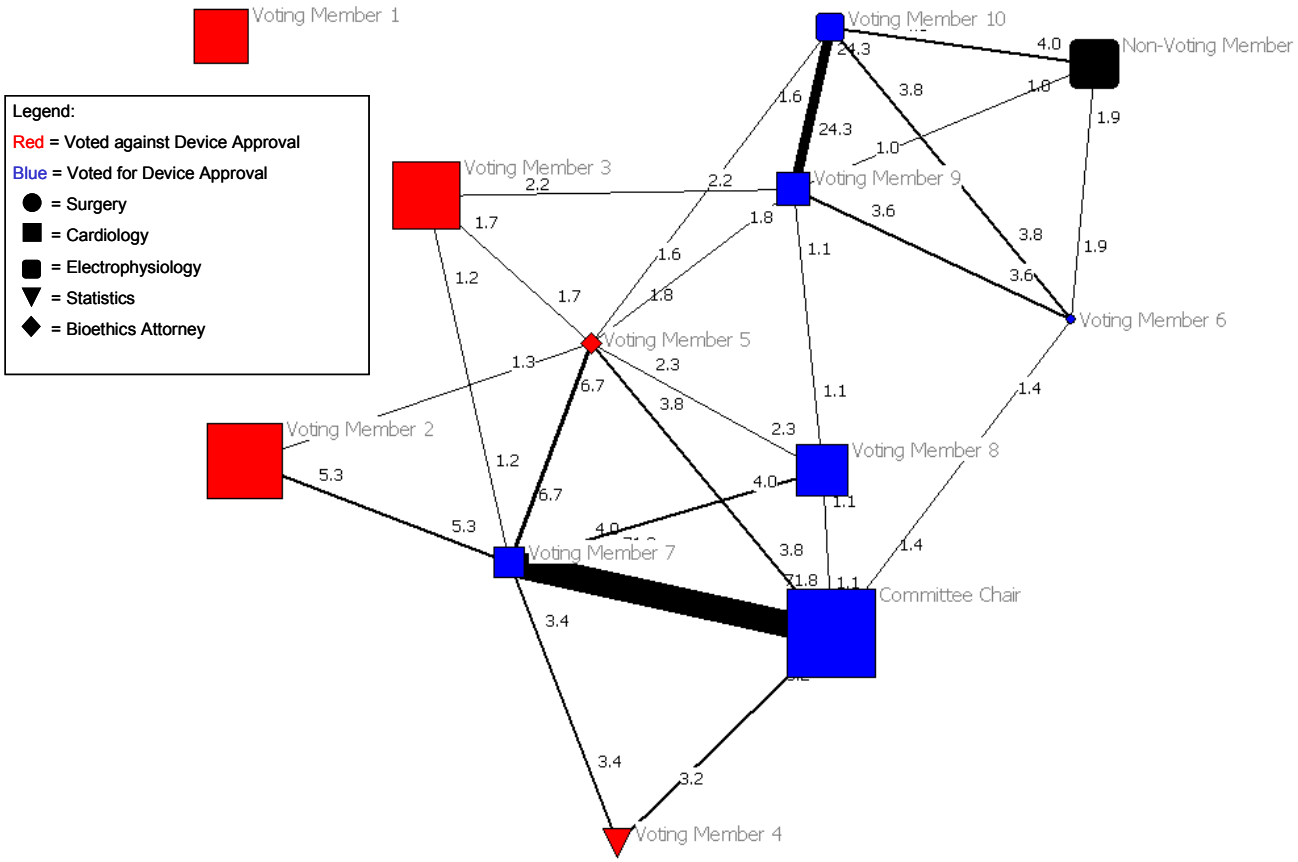


Figure 26: Weighted graph representation of the meeting held on March 5, 2002. Link weights reflect the likelihood that a given edge is due to sharing a topic compared to a background prior distribution. Note that this graph has a similar connectivity pattern to that shown in Figure 13, although it is somewhat denser due to low-likelihood links (e.g., those near 1)

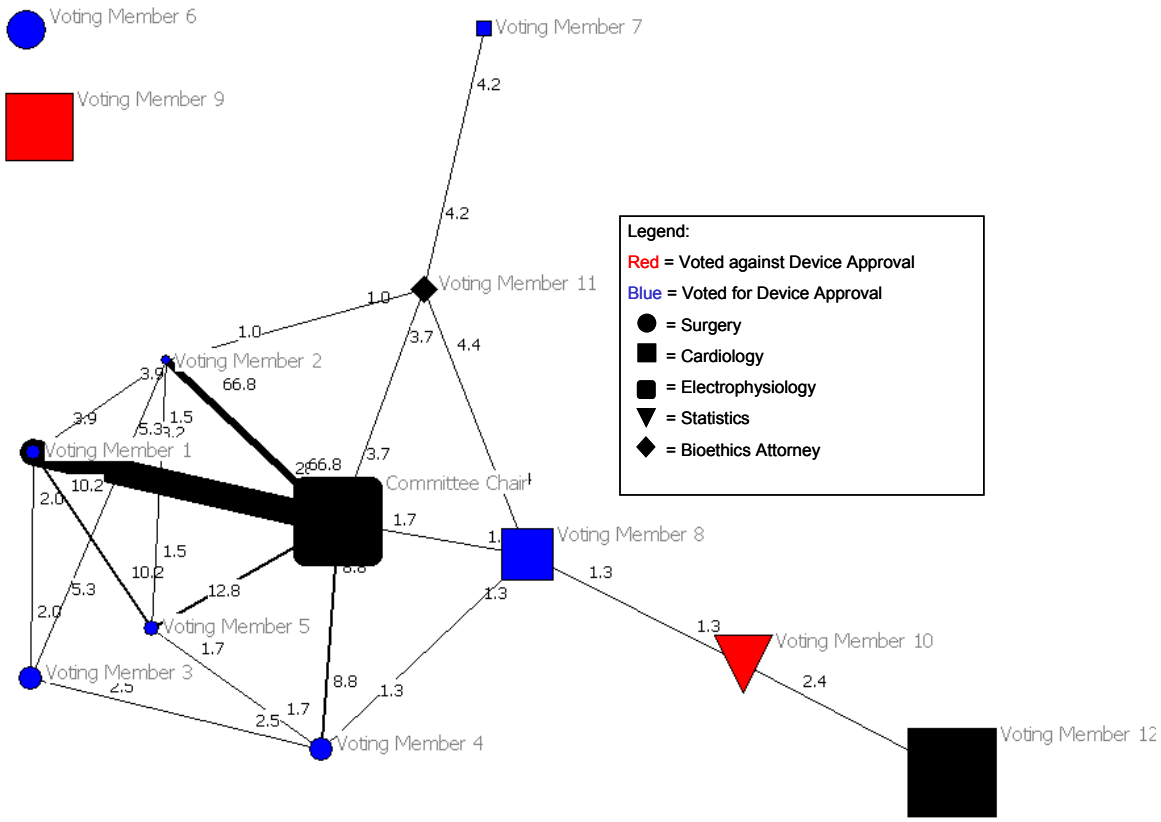


Figure 27: Weighted graph representation of the meeting held on March 5, 2002. Link weights reflect the likelihood that a given edge is due to sharing a topic compared to a background prior distribution. Note that this graph has a similar connectivity pattern to that shown in Figure 15, although it is somewhat denser due to low-likelihood links (e.g., those near 1).

The graphs shown above are qualitatively very similar to those using the Bonferroni cutoff criterion and add apparently more detail. Nevertheless, more precise or explicit representation of uncertainty may be misleading. For example, a link that is 20 times stronger between two speakers does not necessarily imply

an affinity that scales proportionally. Furthermore, a very weak link, with log-likelihood near 1.0, is likely to be due to noise. The above method of weighting graph links may prove useful in future research. Still, our analysis is more concerned with the presence or absence of links than their weights. Therefore, all subsequent results rely on the Bonferroni cutoff criterion. Nevertheless, future work could focus on refining the signal-detection scheme described above.

Comparison across time

Having established a means of grouping voters in a social network, we would now like to be able to include a temporal aspect in the analysis so as to be able to examine patterns of influence. Early attempts to do so focused on the idea that each FDA meeting may be divided into sections that coincide with natural breaks in the meeting. Examples of such include lunch, and coffee breaks. These breaks provide natural stopping points for an analysis. In addition, it is precisely during these breaks that committee members may share information off-the-record that would otherwise remain unshared. Thus comparing pre- and post-break graphs might provide insight into the evolution of committee decisions. All graphs shown in this section use a linkage threshold of 20% with ten topics. Figure 28, Figure 29 and Figure 30 show the social networks of the January 13, 2005 meeting for the amount of time between each break:

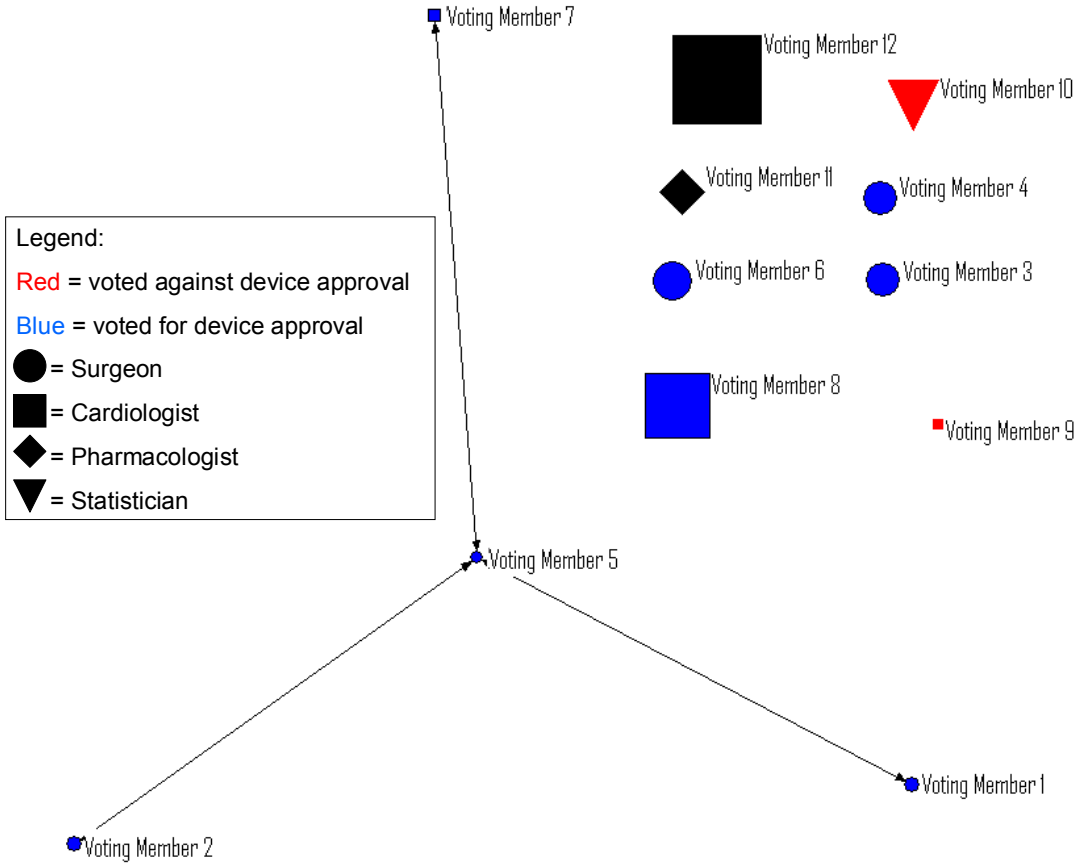


Figure 28: First segment of the January 13, 2005 Circulatory Systems Devices Panel Meeting. At this point in the meeting, voting members had not yet expressed any preferences regarding voting. Rather, committee members were listening to the open public hearing and sponsor presentations. Data include utterances 1-377 of 1671 total utterances.

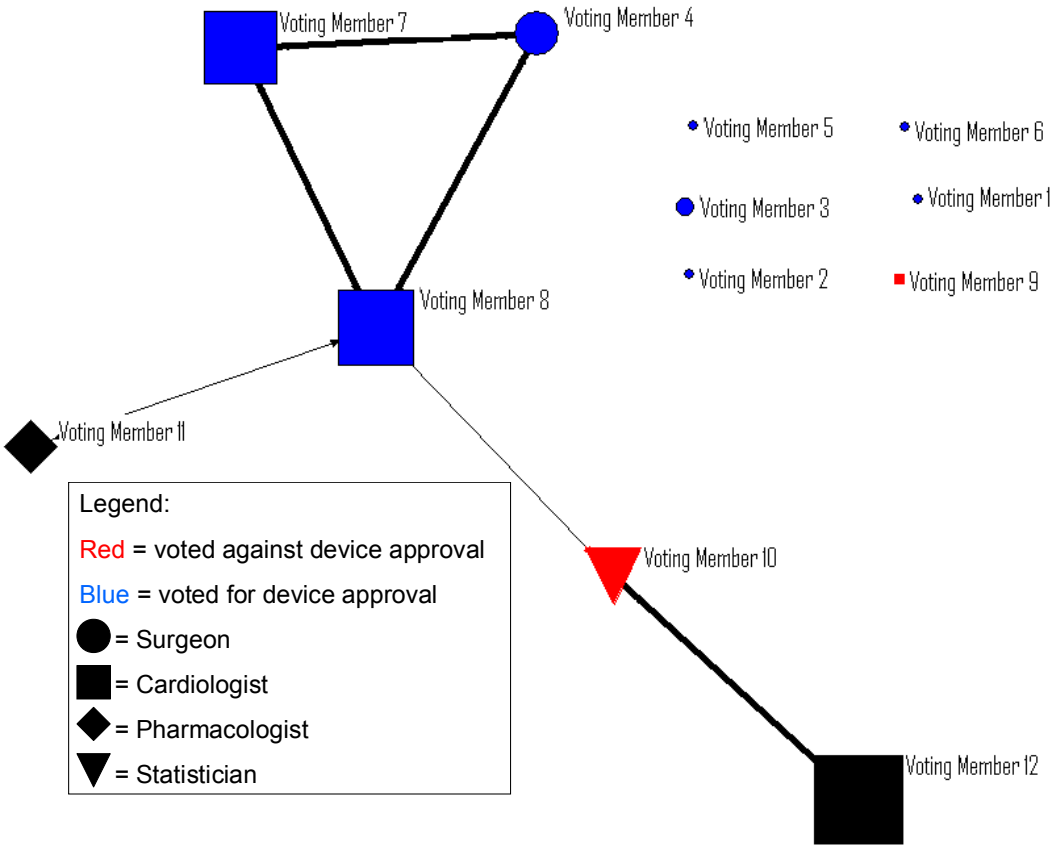


Figure 29: Second segment of the January 13, 2005 Circulatory Systems Devices Panel Meeting. This graph shows that, at this point in the meeting, Voting Members 5, 7, 8, 10, 11 and 12 had begun discussing the statistical elements of the clinical trial design. Five of the six surgeons present have not yet expressed utterances. Data include utterances 378-589 of 1671 total utterances.

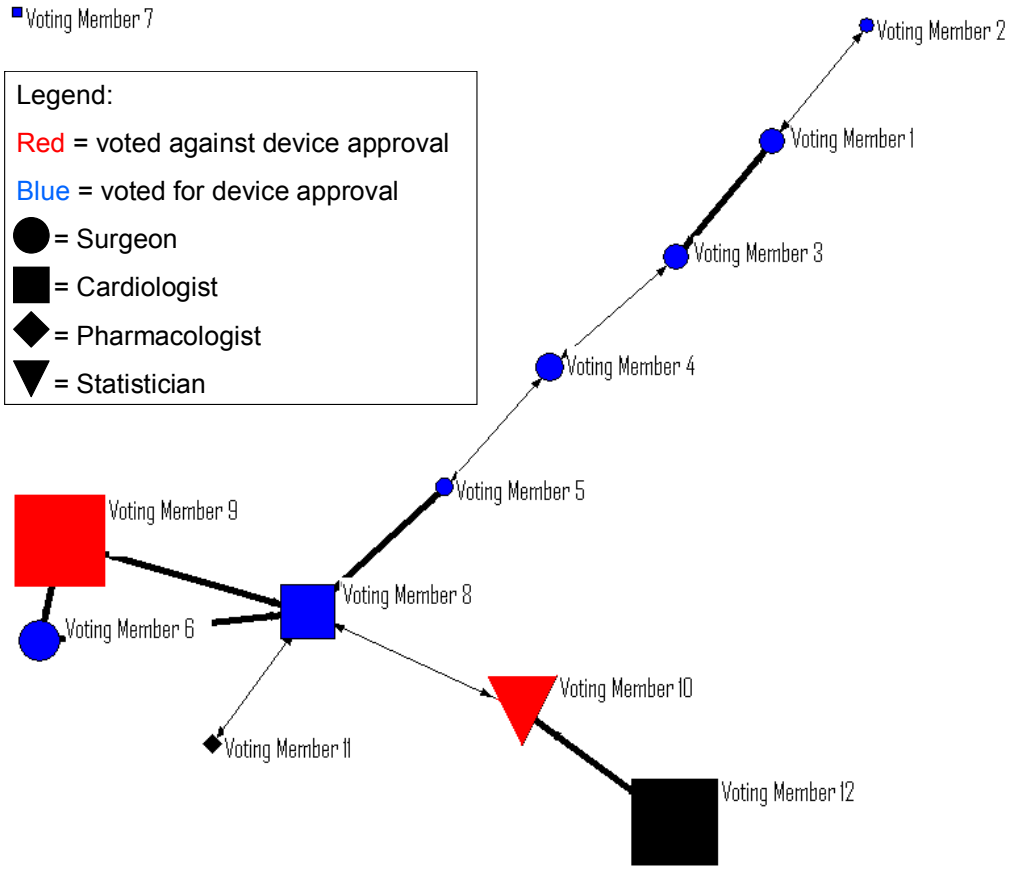


Figure 30: Third, and final, segment of the January 13, 2005 Circulatory Systems Device Panel Meeting. This graph shows that, after lunch, the surgeons in the room, who were previous silent, seemed to align in favor of device approval. Voting Members 8, 9, 10 and 12 seemed to maintain their relative positions between the second and third segments. Data include utterances 590-1671.

The above figures show a small group of voters engaging in a discussion of interest – forming a coalition, as it were – while those who remain silent

eventually come to dominate the voting outcome through strength of numbers. It is particularly interesting that these two groups may be roughly divided by medical specialty, with exchange between representatives of each specialty group having appeared by the third segment.

We may perform a similar analysis on the meeting analyzed previously – i.e., the meeting of the Circulatory Systems Devices Panel of March, 5th, 2002. This meeting is divided into "before lunch" and "after lunch" segments, as shown in Figure 31 and Figure 32.

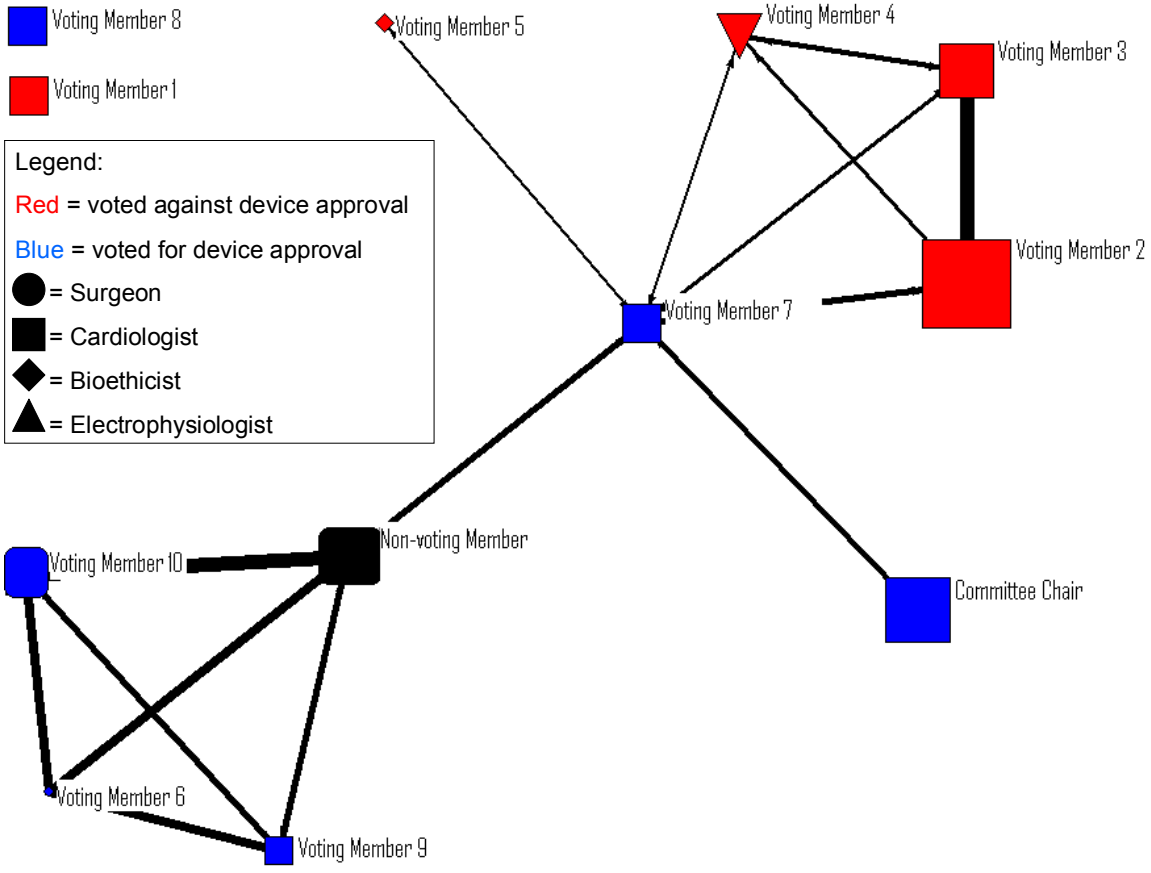


Figure 31: Before-lunch segment of the March 5th, 2002 Circulatory Systems Devices Panel Meeting. This graph shows that, at this point in the

meeting, voting members had largely aligned themselves into blocs that would later vote similarly. Data include utterances 1-703 of 1250 total utterances.

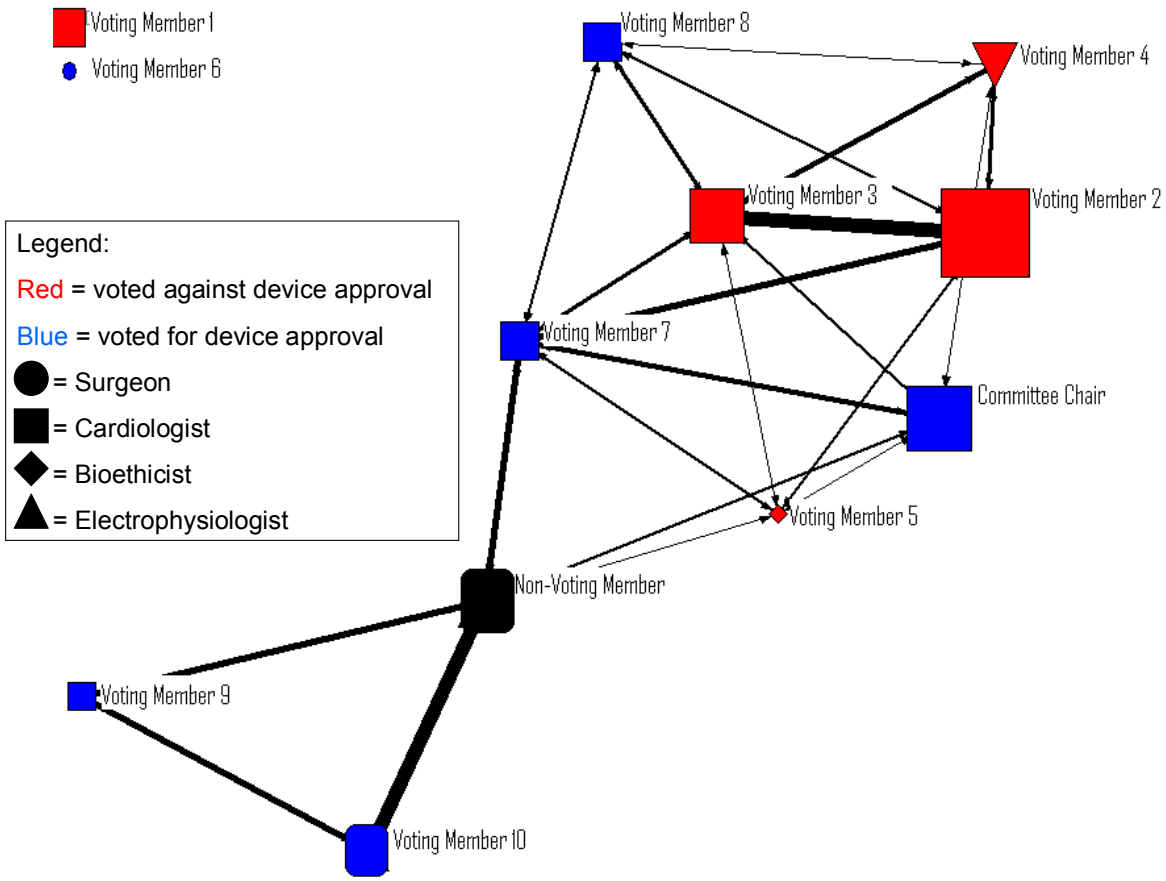


Figure 32: After-lunch segment of the March 5th, 2002 Circulatory Systems Devices Panel Meeting. This graph shows that, by the second half of the meeting, those who would later vote against device approval had become more strongly linked to those who would later support device approval. This pattern perhaps reflects attempts by the approval voters to convince the

non-approval voters to vote differently. Data include utterances 704-1250 of 1250 total utterances.

These graphs indicate a strong grouping by vote prior to lunch, followed by communication across these groups afterwards. Voting seemed to occur along the lines established early in the meeting.

Finally, we examine a meeting held on April 21, 2004. This meeting was originally divided into four parts. Given that the voting members did not speak during the first two quarters of the meeting (leading to a fully disconnected graph), we present only the last two parts of the meeting, displayed in Figure 33 and Figure 34.

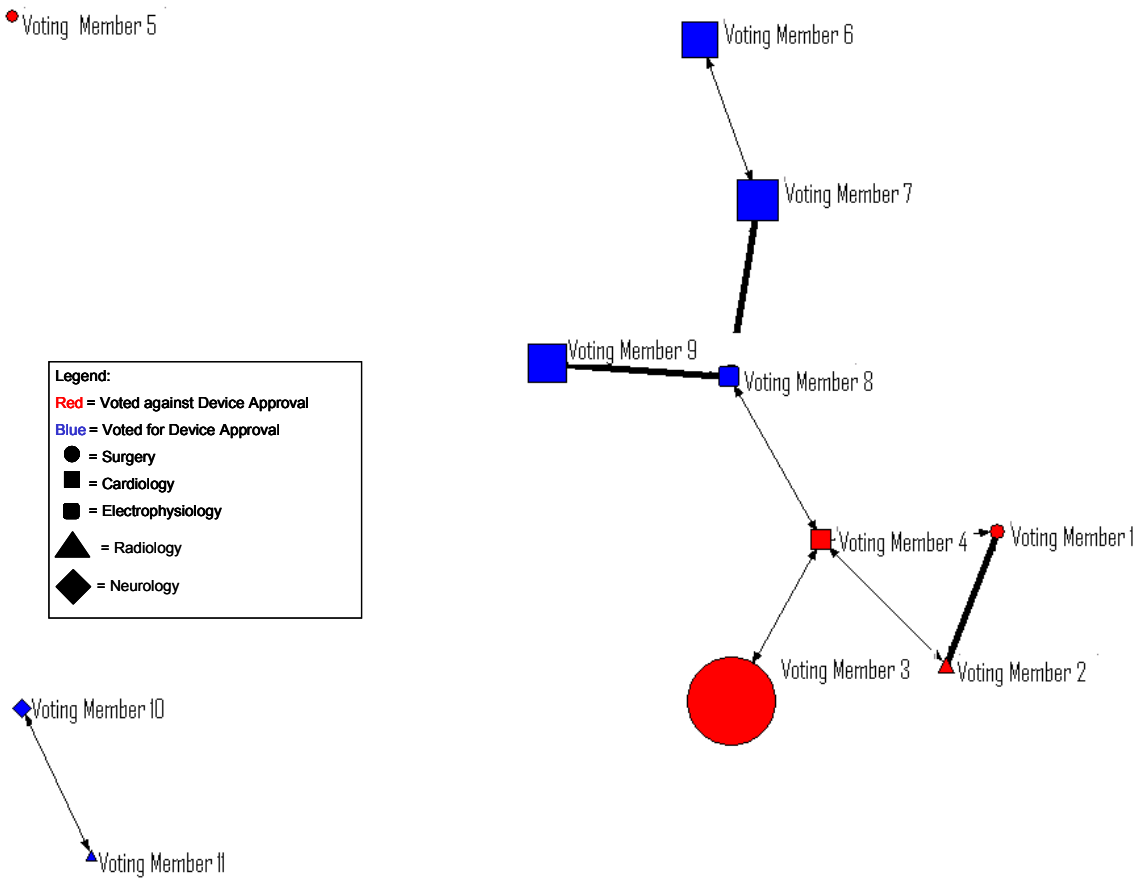


Figure 33: Before-lunch segment of the April 21st, 2004 Circulatory Systems Devices Panel Meeting. This graph shows well-defined coalitions having been formed relatively early in the meeting. It is interesting that voting patterns seem to largely respect the boundaries of particular medical specialties (i.e., surgeons vs. cardiologists). Data include utterances 399-876 of 1822 total utterances.

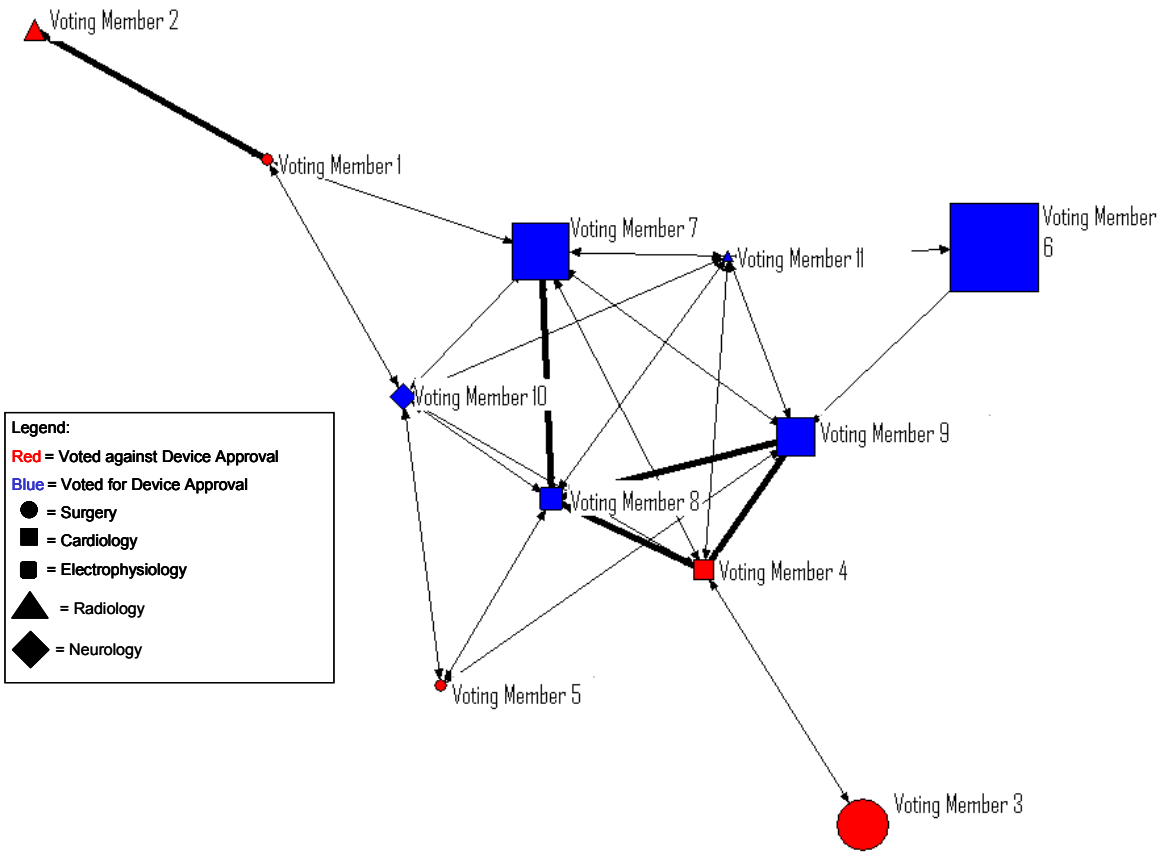


Figure 34: After-lunch segment of the April 21st, 2004 Circulatory Systems Devices Panel Meeting. This graph shows that the well-defined coalitions of the before-lunch segment have broken down – particularly the anti-device coalition. This may well be due to attemtps by members of one coalition to influence the other, leading to cross-coalition dialogue.. Data include utterances 877-1822 of 1822 total utterances.

The first meeting segment shows the formation of two coalitions that ultimately voted oppositely. It is interesting that the pro-approval coalition is composed largely of cardiologists, whereas the anti-approval coalition is composed largely of non-cardiologists. Furthermore, the bridging members, Voting Members 4 and 8 were outliers within their own group. Both served as chairs of other meetings, and are therefore perhaps more likely to listen broadly and to work to achieve consensus among panel members. The second meeting segment shows the breakdown and fragmentation of the anti-approval coalition and the consolidation of the pro-approval coalition prior to voting which may again indicate that, later in the meeting, attempts at dialogue across groups occurred but did not achieve consensus in this case. Because we do not know an individual's preference midway through the meeting, we cannot tell if this effect holds during meetings that did reach consensus.

Extraction of time dynamics using the above method is not generalizable across meetings. The reason for this is that the locations of the lunch and coffee breaks are not always timed to coincide with speech from voting members. In many meetings, the lunch break occurs before any panel member has an opportunity to speak. It is difficult to tell, from this representation, how influence passes in these committee meetings. Furthermore, a representation that separates post-break from pre-break implicitly assumes that no words spoken before the break carry over – this is clearly incorrect. Finally, there may be significant dynamics that occur on a shorter time-scale than depicted above. Although future work could focus on further developing this technique using methods such as Dynamic Network Analysis (Carley 2003), this thesis presents a different method of incorporating time into the analysis. This will be presented below.

Directed Graphs

The conversation analysis literature in sociology (e.g., Gibson 2008) notes that, within small groups, influence is often linked to capacity to affect a topic shift. This is because of the linear nature of speech – two people cannot typically speak at the same time if both are to be understood. Speaking order is therefore related to agenda control. For example, a more influential speaker may change the subject, whereas a less influential speaker will remain on the subject introduced by the higher-status speaker.

Given an infrastructure for examining topic overlap among speakers, we can take advantage of the temporal aspect of our data to develop insights about topic changing as follows:

Consider a sample, s, from the posterior distribution of the AT model. Within this sample, choose a pair of speakers, x_1 and x_2, and a topic z. Given that utterances are temporally ordered, this defines two separate time-series. Figure 35 shows two time series for two different speakers in the meeting held on March 4, 2002.

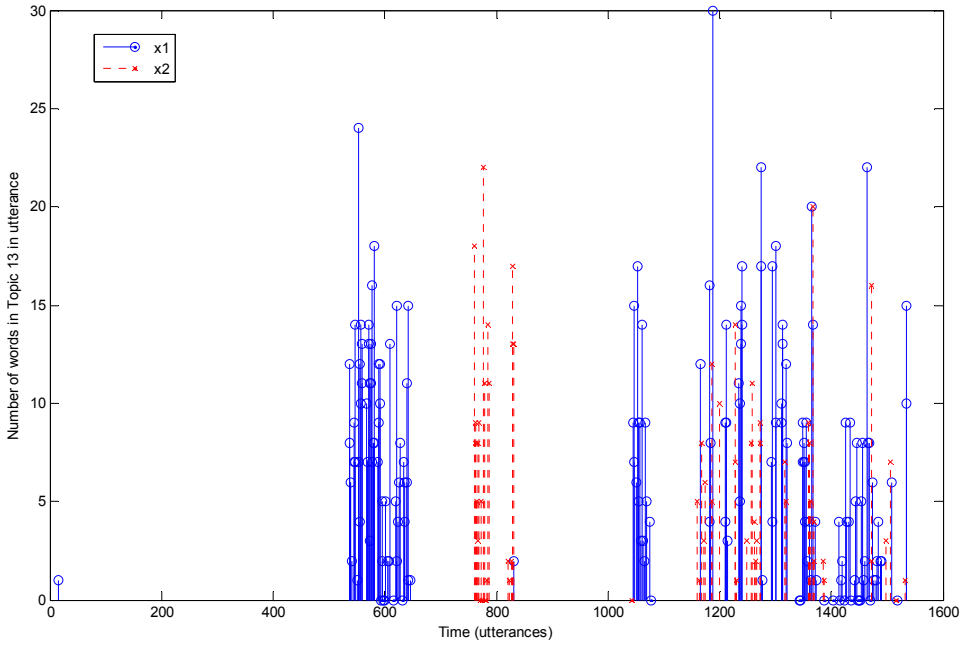


Figure 35: Time series for two speakers on topic #13 during the meeting held on January 13, 2005.

This chart clearly shows that x_1 speaks about topic z before x_2 does. Based on this, we can say that x_1 *leads* x_2. These time series can be used to generate the *topic-specific cross correlation* for speakers x_1 and x_2, in topic z:

$$(f_{i,z}^{s} * f_{j,z}^{s})[\delta] = \sum_{d=-\infty}^{\infty} f_{i,z}^{s}{}^{*}[d] f_{j,z}^{s}[\delta + d] \qquad \textbf{(11)}$$

where $f_{i,t}^{s}(d)$ is the number of words spoken by author i and assigned to topic z in document d, in sample s. The cross-correlation function for the data shown in Figure 35 is shown in Figure 36.

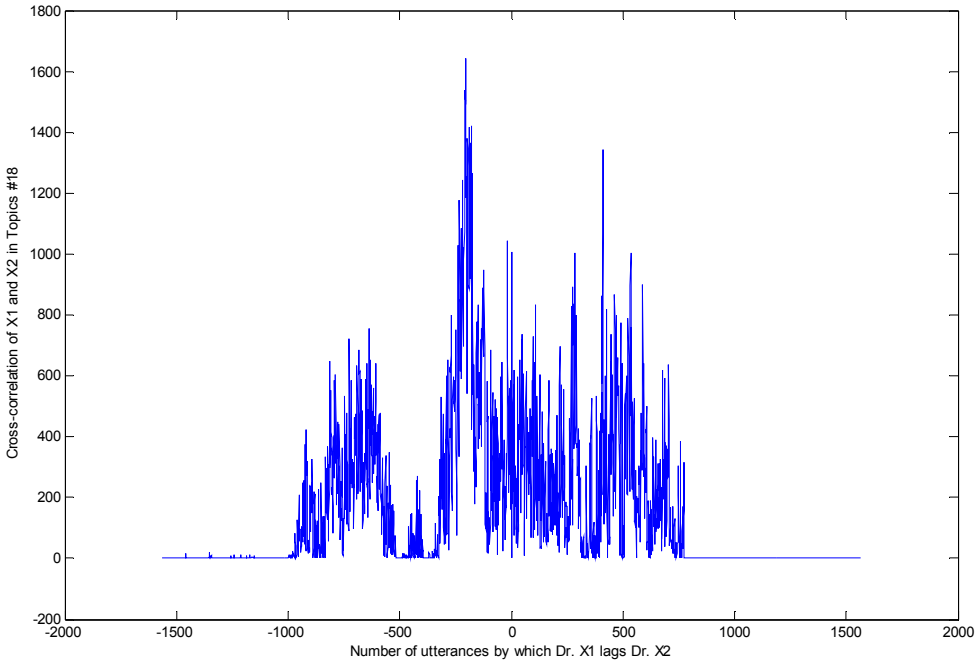


Figure 36: Cross-correlation of the two time series shown in Figure 35.

Figure 36 clearly shows that the maximum value of the cross-correlation function is less than zero. This is a quantitative indication that x_2 *lags* x_1. The location of this peak is described by the expression $m_1 = \arg\max_{\delta}(f^s_{i,z} * f^s_{j,z})[\delta]$. In principle, there may be multiple peaks in the cross-correlation function, as in Figure 37.

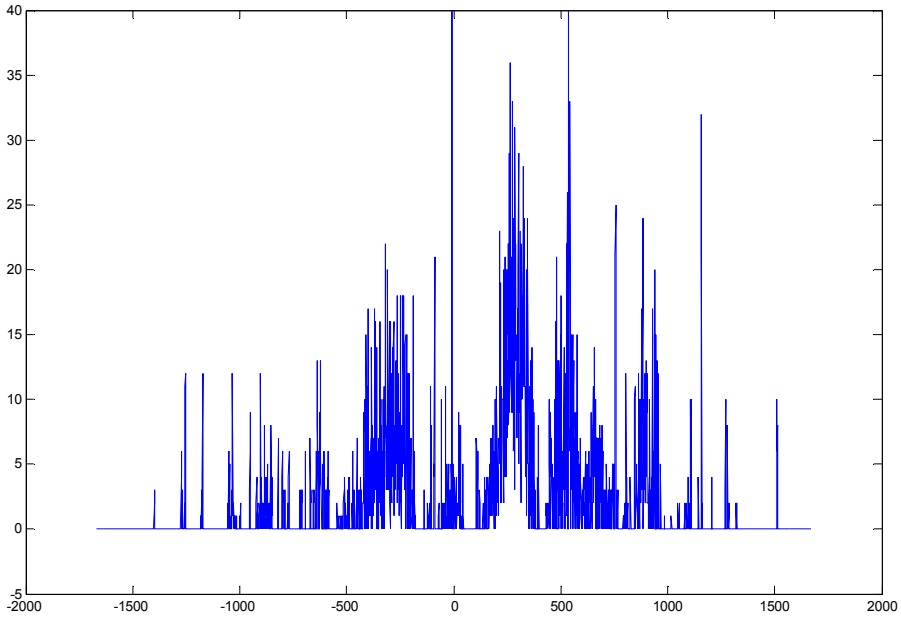


Figure 37: A cross-correlation function with two peaks, representing two speakers who are equally involved in leading conversation on this topic.

For each sample, s, from the AT Model's posterior distribution, we examine the cross-correlation function for each author pair, $\{x_i, x_j\}$, in topic z. Let there be k peaks in the cross-correlation function. For each peak, if $m_k > 0$, we say that author i *lags* author j in topic z, at point m_k (i.e., $l^s_{i,j,z,m_k} = 1$). Similarly, we say that author i *leads* author j in topic z at point m_k (i.e., $l^s_{i,j,z,m_k} = -1$) if $m_k < 0$. Otherwise, $l^s_{i,j,z,m_k} = 0$. For each sample, s, we define the *polarity* of authors i and j in topic z to be the median of the $l^s_{i,j,zt}$.

$$p^s_{i,j,t} = median(l^s_{i,j,z})$$

If most of the peaks in the cross-correlation function are greater than zero, then the polarity = 1; if most of the peaks are less than zero, then the polarity = -1; otherwise, the polarity = 0.

We are particularly interested in the topic polarities for author-pairs who are linked in the graph methodology outlined above – i.e., where $\Delta_{i,j} = \Delta_{j,i} = 1$. Using the polarity values defined above, we are interested in determining directionality in $\mathbf{\Delta}$. For each sample, s, we define the *direction* of $e_{i,j}$ in sample s as:

$$d^s(e_{i,j}) = \sum_{t=1}^{T} (p^s_{i,j,t} * P^s(Z = z_i \mid x_i) * P^s(Z = z_i \mid x_j)) \qquad \textbf{(12)}$$

This expression weights each topic polarity by its importance in the joint probability distribution between x_i and x_j, and is constrained to be between -1 and 1 by definition. The set of 200 $d^s(e_{i,j})$ defines a distribution, three types of which are shown below:

The *net edge direction,* $d(e_{i,j})$ is determined by partition of the unit interval into three equal segments. In particular, we examine the proportion of the $d^s(e_{i,j})$ that are greater than 0. If more than 66% of the $d^s(e_{i,j}) > 0$ then $d(e_{i,j}) = 1$ (the arrow points from j to i) – an empirical example of this case is shown in Figure 38.

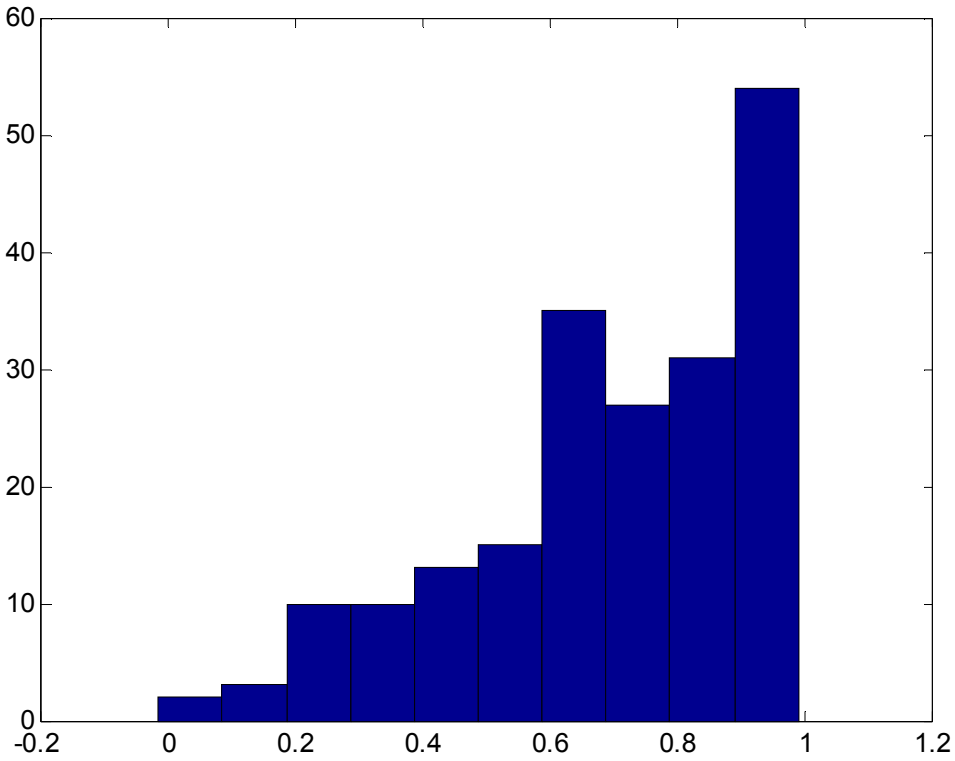


Figure 38: Edge direction distribution for two speakers, one of who clearly leads the other. Both speakers were voting members in the meeting held on January 13, 2005.

If less than 33% of $d^s(e_{i,j}) > 0$ then $d(e_{i,j}) = -1$ (the arrow points from i to j) – an empirical example of this case is shown in Figure 39.

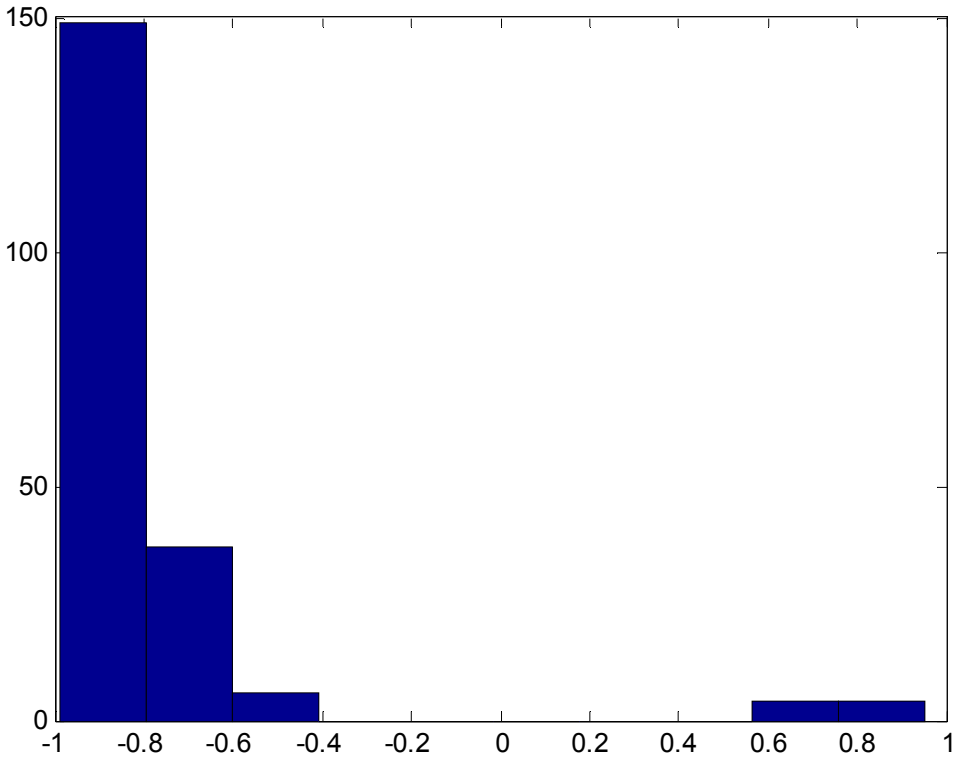


Figure 39: Edge direction distribution for two speakers, one of whom clearly lags the other. Both speakers were voting members in the meeting held on January 13, 2005.

Otherwise, $d(e_{i,j}) = 0$ (the arrow is bidirectional) – an empirical example of this case is shown in Figure 40.

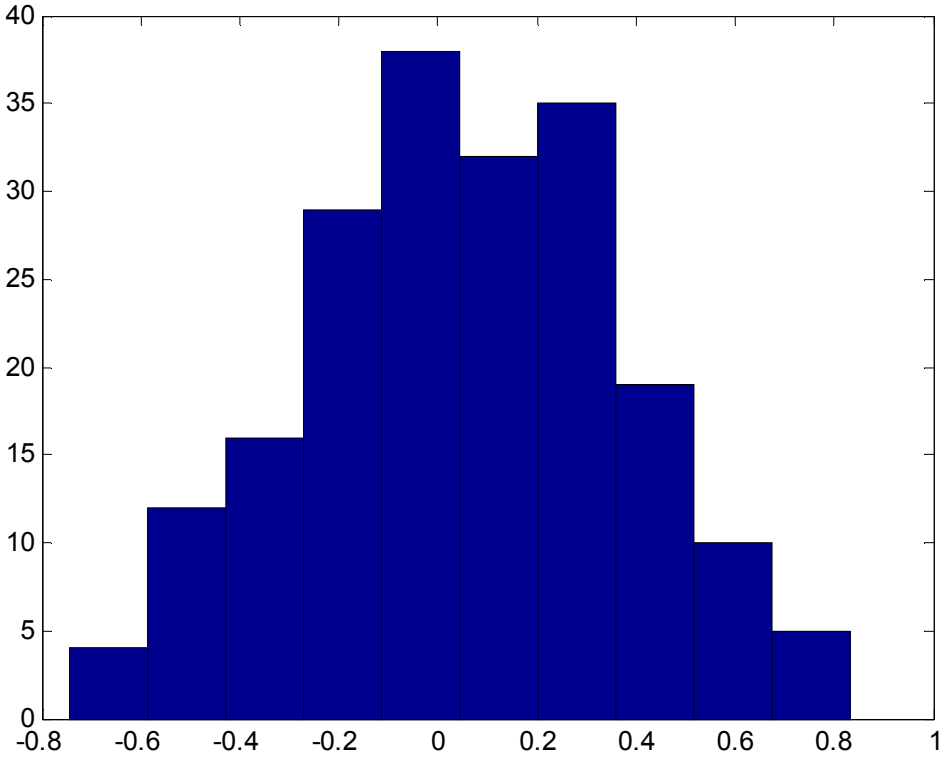


Figure 40: Edge direction distribution for two speakers, neither of whom clearly lags the other. Both speakers were voting members in the meeting held on January 13, 2005.

The result is a directed network, examples of which are seen in Figure 41 and Figure 42.

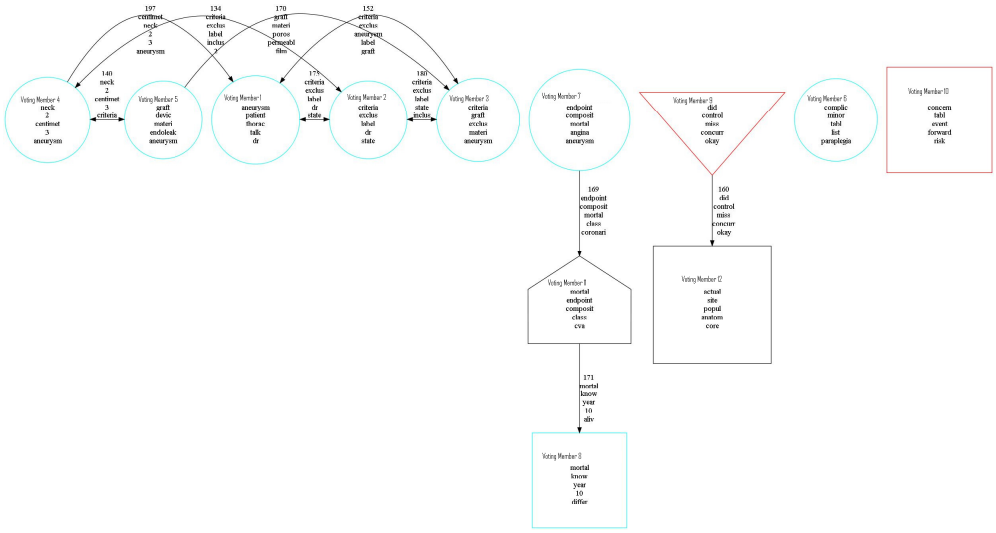


Figure 41: Directed network representation of the FDA Circulatory Systems Advisory Panel meeting held on January 13, 2005. Node size increases with the number of words spoken by that author; node shape represents medical specialty. Non-approval votes are red; approval votes are blue; non-voters are black. Each speaker's top five words are listed, as is each edge's link frequency. This diagram is generated using the dot algorithm (Gansner and North 1999).

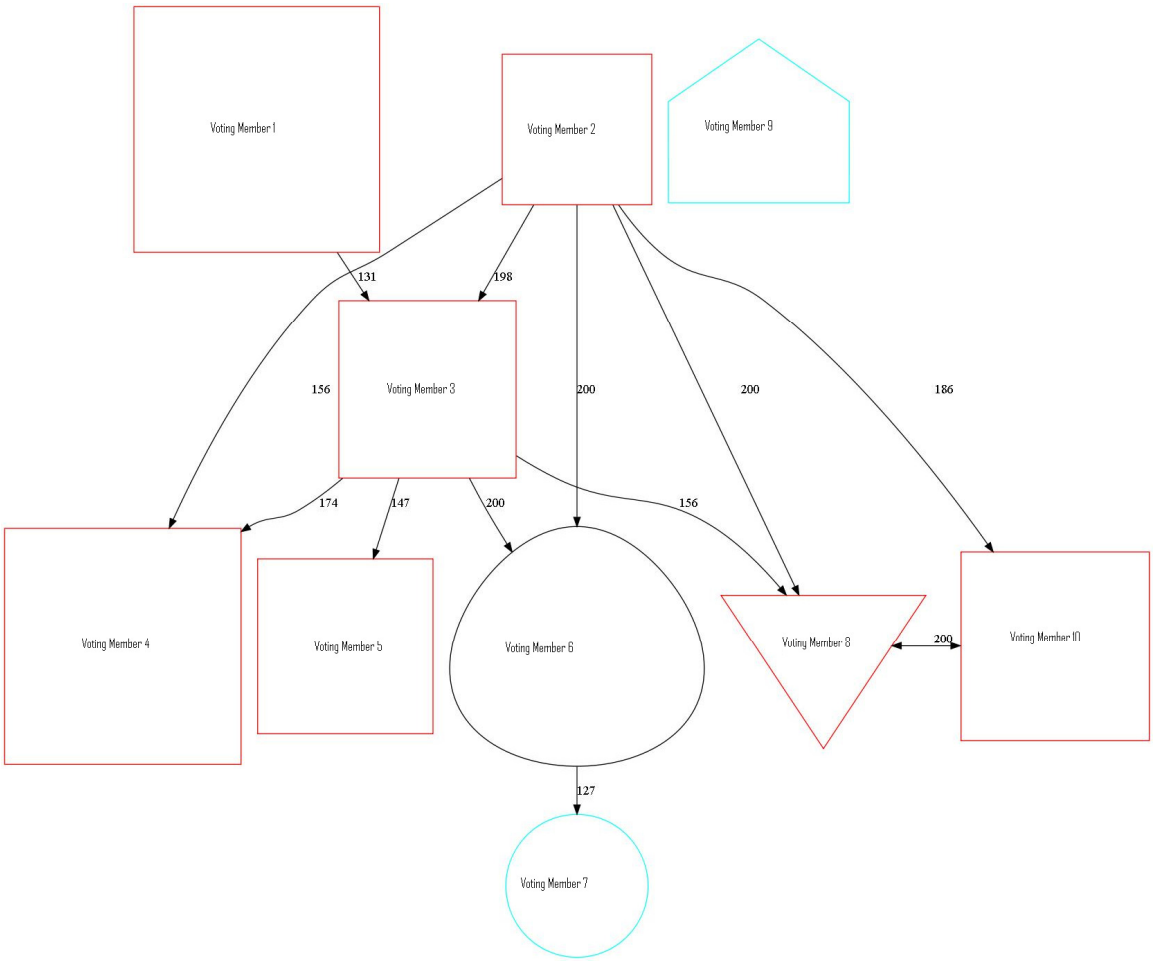


Figure 42: Directed network representation of the FDA Circulatory Systems Advisory Panel meeting held on July 9, 2001. Node size increases with the number of words spoken by that author; node shape represents medical specialty. Non-approval votes are red; approval votes are blue; non-voters are black. This diagram is generated using the dot algorithm (Gansner and North 1999).

These directed graphs address all identified limitations of the LSA approach while simultaneously providing a computational platform for the analysis of

communication patterns in technical expert committee meetings. The next chapter focuses on analysis of these graphs and presents results regarding decision-making on FDA panels.

C h a p t e r 5

RESULTS AND IMPLICATIONS

אמר אביי ולטעמיה דרב בא אחד ואמר לה הרי את מקודשת לי מעכשיו ולאחר ל' יום ובא אחר ואמר לה הרי את מקודשת לי מעכשיו ולאחר עשרים יום ובא אחר ואמר לה הרי את מקודשת לי מעכשיו ולאחר עשרה ימים מראשון ומאחרון צריכה גט מאמצעי אינה צריכה גט מה נפשך אי תנאה הואי דקמא קידושי דהנך לאו קידושי אי חזרה הואי דבתרא קידושי דהנך לאו קידושי פשיטא מהו דתימא האי לישנא משמע תנאה ומשמע חזרה ותיבעי גיטא מכל חד וחד קמ"ל

Abaye said: According to Rav, if [a man] came and said to [a woman]: "Behold, you are betrothed to me from now and after thirty days,"and then another man came and said to [the same woman]: "Behold, you are betrothed to me from now and after twenty days," and then another man came and said to her: "Behold, you are betrothed to me from now and after ten days," she requires a divorce from the first and from the last, but from the intermediate she does not require a divorce. Whatever you consider, if [each man's statement] is a stipulation, the first man's betrothal [is valid]. If, [each man's statement] is a retraction, the last man's betrothal [is valid]; the other mens' betrothals are not. This is obvious; [but] you might have interpreted [Rav to mean that] this language can carry the meaning [of] a stipulation, and can [also] carry the meaning [of] a retraction. Thus [the woman] would require a divorce from each and every one. [Abaye] informs us. –Babylonian Talmud, Kiddushin, 59b-60a, trans. Hebrew and Aramaic

"It is not that the rabbis went to the depth of peoples' minds that they either all mean a stipulation or they all mean a retraction. Rather, the implication of the words was uncertain to them. Therefore, even if one of them says that he meant one thing, and the other says he meant the opposite, we only pay heed to the primary indication and law of the language. For the intention of the one who made the betrothal isn't known to the witnesses except from the language and what it indicates." – Rabbi Shlomo ben Aderet, b. 1235 - d. 1310, Chiddushei HaRashba. trans. Hebrew, Rabbi B. Ganz, on socially determined limits of verbal ambiguity.

Sources of Influence on FDA Panels

One of the major goals of this work is to attempt to identify potential flows of communication on FDA panels, and their implications for committee behavior. In order to better understand these, we first examine the backgrounds of individual panel members. The following information has been collected for each panel member:

1. Gender
2. Race
3. Medical Specialty
4. Age (number of years since doctoral-level degree granted)
5. h-Index[8]

The first four attributes were collected using a combination of Google searches and information stored at http://www.vitals.com; whereas h-Index for a given panel member in a given year was provided by the ISI Web of Science.

With the exception of medical specialty, all of the variables listed above fall into the category of "attribute-based status characteristics", as defined in the "small-groups" strand of literature in sociology (e.g., as represented in Berger et al. 1972). This body of literature predicts that these status characteristics might be associated with different voting behaviors. One sort of behavior that we might see on FDA panels is "air-time" – i.e., the amount of time that a given speaker speaks. Research in social psychology has shown that perceived influence is

[8] H-Index is a metric of academic prestige associated with journal citation behavior. "*A scientist has index h if h of [his/her] N_p papers have at least h citations each, and the other (N_p − h) papers have at most h citations each.*" (Hirsch 2005). Deeper analysis of the impact of other demographic variables on h-index (cf. the analysis performed in Kelly & Jennions 2006) is shown in Appendix 2.

associated with air-time (Bottger 1984). This is to be contrasted with actual influence, which Bottger finds is associated with problem-solving expertise. Results shown in Table 6 indicate that several variables have a significant effect on air-time:

Table 6: 4-way ANOVA showing the effects of Gender, Medical Specialty, h-Index, and Age on air-time for our sample of 37 meetings. In this analysis, air-time has been normalized and a logit transform has been applied to enable comparisons across meetings. When race is included as an explanatory variable, it fails to reach significance (p=0.20), suggesting no identifiable effect of race. Medical Specialty captures most of the variance in air-time, followed by h-Index, gender and age.

Variable	Sum of Squares	Degrees of Freedom	Mean Squares	F	p-value
Gender	5.81	1	5.81	14.24	0.0002
Medical Specialty	9.69	7	1.38	3.39	0.0016
h-Index	7.19	1	7.19	17.64	<0.0001
Age	2.61	1	2.61	6.40	0.012
Error	137.78	338	0.41		

Total	168.44	348

Independent Tukey Honestly-Significant Difference (HSD) tests of multiple comparisons show that women use significantly more air-time than do men; however, this effect does not exist for the subset of 17 meetings in which a voting minority existed. Table 7 shows the same ANOVA analysis on this subset of meetings:

Table 7: 4-way ANOVA showing the effects of Gender, Medical Specialty, h-Index, and Age on air-time for the subset of 17 meetings in which there was a minority. In this analysis, air-time has been normalized and a logit transform has been applied to enable comparisons across meetings. Here, most of the variance is captured by h-Index followed by Medical Specialty and age. Gender fails to reach significance as an explanatory variable.

Variable	Sum of Squares	Degrees of Freedom	Mean Squares	F	p-value
Gender	0.0022	1	0.0022	0.93	0.34
Medical Specialty	0.039	6	0.0065	2.73	0.015
h-Index	0.043	1	0.043	18.22	<0.0001

Age	0.013	1	0.013	5.28	0.023
Error	137.78	338	0.41		
Total	168.44	348			

Empirical Finding 1: Gender, Medical Specialty, h-Index, and Age are all significant variables associated with a panel member's air-time. Women tend to have more air-time than men do, although this effect is not visible in meetings with voting differences.

Bottger found that, in the most effective teams, air-time and actual influence covary. We might therefore expect that, in meetings where there is not consensus, members of the voting majority would tend to have a higher air-time than do members of the voting minority. Figure 43 shows that this is not the case.

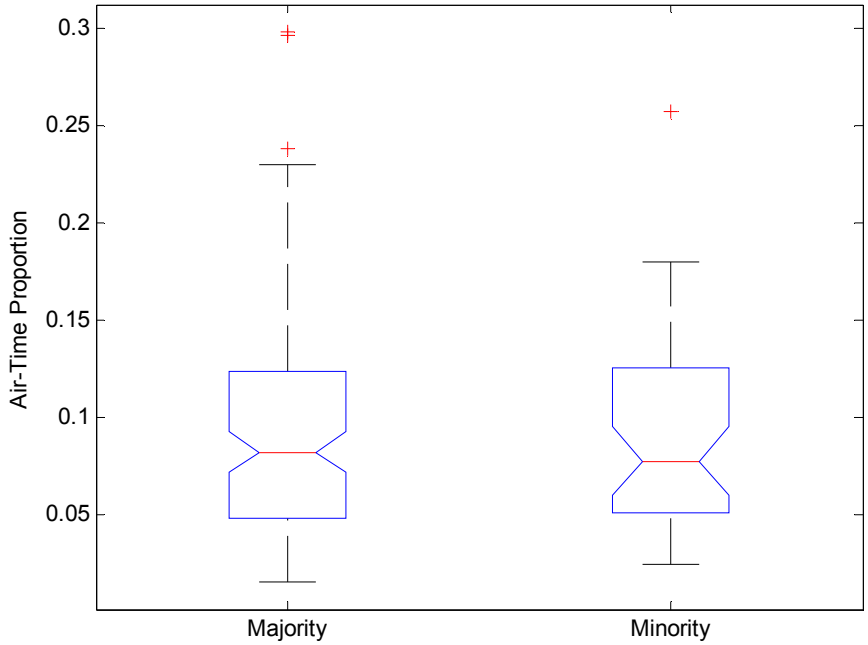


Figure 43: A Kruskal-Wallis test shows no significant difference between the air-time proportions of majority and minority voters (p=0.86) for the 17 meetings in which a split-vote existed.

Empirical Finding 2: There is no observably significant effect between vote and air-time.

Finally, we might examine the impact of many of the status characteristics outlined above on voting behavior. Although vote is a dichotomous variable, and therefore does not meet the ANOVA assumptions, one may argue that, with a sufficiently large number of datapoints, ANOVA still provides useful results with a sufficiently large number of degrees of freedom for error (Lunney 1970; see Table 8).

Table 8: 4-way ANOVA showing the effects of Gender, Medical Specialty, h-Index and Age on voting outcome for the 17 meetings in which there was a voting minority. In this analysis, voting outcome is a dichotomous variable, thereby violating the ANOVA assumptions. Only gender has a significant effect on voting outcome.

Variable	Sum of Squares	Degrees of Freedom	Mean Squares	F	p-value
Gender	1.442	1	1.442	7.76	0.006
Medical Specialty	0.5096	6	0.085	0.46	0.84
h-Index	0.039	1	0.039	0.21	0.65
Age	0.097	1	0.097	0.52	0.47
Race	0.51	3	0.17	0.91	0.42
Error	29.72	160	0.19		
Total	32.31	172			

The above table shows no significant effect of any of the variables tested on voting outcome, with the exception of gender. Table 9 shows an independent

analysis of the effect of gender on voting outcome, with the result that women are more frequently in the voting majority than men are.

Table 9: A chi-square test examining the impact of gender on voting outcome for the 17 meetings in which a minority existed shows a significant result (χ^2=8.29;dof=1;p=0.0040) with women more likely to be in the majority.

	Majority	Minority	TOTAL
Male	100	41	141
Female	33	2	35
TOTAL	133	43	176

Empirical Finding 3: There is no observably significant effect of medical specialty, h-Index, age or race on voting behavior. Women are more likely to be in the voting majority than men are.

Medical Specialty as an Organizing Factor

We now turn to the role of medical specialty as an organizing factor on the FDA Circulatory Systems Advisory Panel. Unlike the other characteristics discussed in (Berger et al. 1972), medical specialty is typically not associated with status in the sociology literature. Indeed, Table 10 shows that medical specialty alone is not a strong predictor of voting behavior:

Table 10: There is no significant relation between medical specialty and voting behavior (χ^2=4.29;dof=8;p=0.83)

	Surgeon	Cardio-logist	Electrophysio-logist	Statistician	Other	TOTAL
Abstention	2	2	1	1	2	8
Voting Minority	15	16	6	2	4	43
Voting Majority	33	53	15	12	20	133
TOTAL	50	71	22	15	26	184

Empirical Finding 4: There is no observably significant effect between medical specialty and vote.

We have already noted the statistically significant role of medical specialty as a control variable when measuring air time. Figure 44 shows a boxplot for the four most strongly-represented specialties on the panel: surgeons, cardiologists, electrophysiologists and statisticians. Visual inspection shows that surgeons and electrophysiologists speak less frequently than do cardiologists and statisticians. A Tukey HSD test for multiple comparisons shows that surgeons speak less frequently than do cardiologists and statisticians, and that cardiologists speak more frequently than do electrophysiologists.

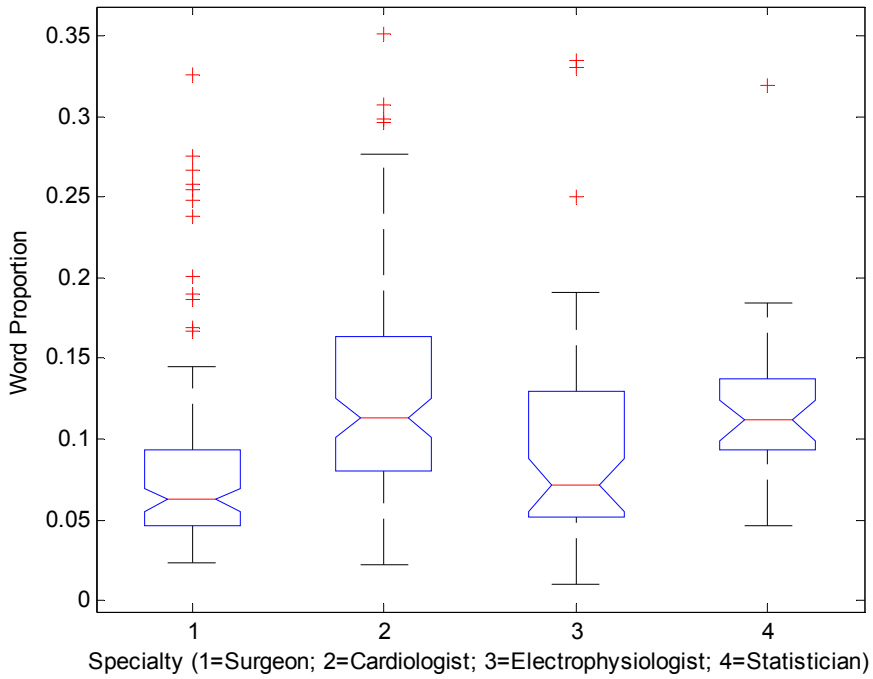


Figure 44: Box plots for the four most strongly-represented specialties. Note that more "clinical" specialties (surgeons and electrophysiologists) tend to speak less than the more "medical" specialties (cardiologists and statisticians).

The mediating effect of medical specialty is perhaps most apparent when it is examined using the graph-based methodology outlined in chapter 3. We have already noted some graphs where individual voters tend to group by medical specialty. This is because, in these meetings, members of the same specialty use common terminology. This strongly suggests that, in these meetings, the subject of discussion is mediated by the specialties present. We would like to formalize this intuition in order to determine if it is a phenomenon that is widespread across meetings:

Consider a graph, **Δ**, generated by the method outlined in chapter 4. One such graph may be generated for each of the 37 meetings that we analyze. We would like to be able to determine, on a given graph, how likely members of the same medical specialty are to be linked to one another. Suppose that graph **Δ** has n edges, m of which connect a pair of speakers who have the same medical specialty. We may therefore define *specialty cohesion* as m/n – the proportion of edges in graph **Δ** connecting members of the same medical specialty. A high specialty cohesion might indicate that members of the same medical specialty are more likely to link than are members of different medical specialties – on the other hand, it might just indicate that the meeting is homogenous – if there is very little diversity on a panel, then we might expect cohesion to be high by definition. We would therefore prefer to compare the observed specialty cohesion to the cohesion of graphs that have similar properties to **Δ.** We can do this by examining *specialty cohesion percentile*: For each graph, **Δ**, representing a meeting, 1000 random graphs are generated having a number of nodes, and a graph density, equal to those found in **Δ.** Each node is similarly assigned a medical specialty as in **Δ**. Specialty cohesion is calculated for each of these random graphs, generating a meeting-specific distribution. Specialty cohesion percentile is defined as the proportion of the resultant graphs that have lower specialty cohesion than **Δ.**

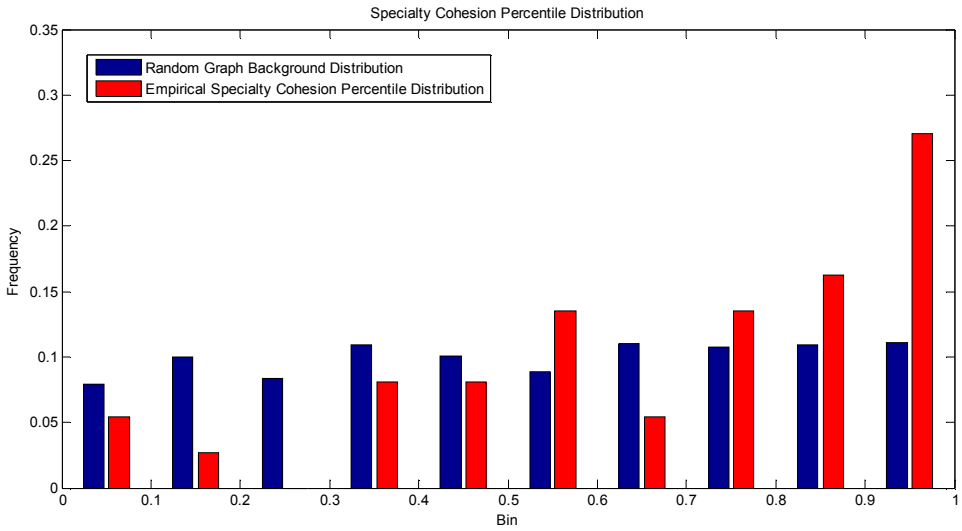


Figure 45: Histogram of Specialty Cohesion Percentiles for the 37 meetings in our sample. The empirical specialty cohesion percentile distribution's cumulative distribution function is significantly less than that of the background distribution (one-sided Kolmogorov-Smirnov test; p=0.0045) indicating that the empirical distribution has more probability density concentrated near unity and away from zero.

Figure 45 shows the empirical distribution of specialty cohesion percentiles for the 37 meetings analyzed (in red). This is contrasted with the specialty cohesion percentile distribution for 1000 random graphs – a uniform distribution. We may see, by inspection, that the empirical specialty cohesion percentile distribution has a right skew – i.e., probability mass is concentrated near 1 and away from 0. This suggests that specialties are more likely to group together than we might expect under conditions of chance. A Kolmogorov-Smirnov test for equality of distributions finds that the empirical cumulative distribution function (CDF) is

significantly less than the uniform background CDF (p=0.0045), indicating that the skew shown in Figure 45 is statistically significant. Plots of the CDFs are shown in Figure 46.

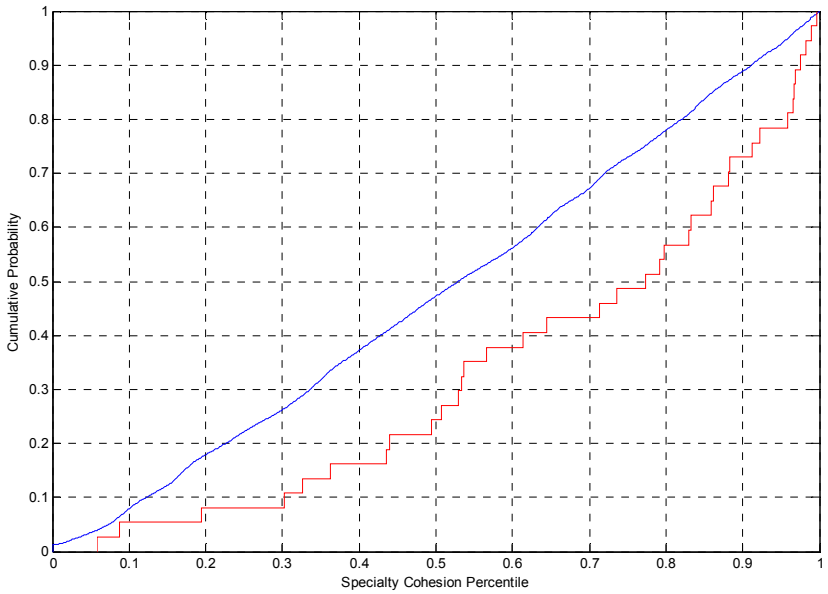


Figure 46: Cumulative Distribution Plot of Specialty Cohesion Percentiles for the 37 meetings in our sample. The empirical specialty cohesion percentile distribution's cumulative distribution function is significantly less than that of the background distribution (one-sided Kolmogorov-Smirnov test; p=0.0045) indicating that the empirical distribution has more probability density concentrated near unity and away from zero.

These results provide support for the notion that members of the same medical specialty tend to preferentially link to one another, but not in a way that totally precludes links to other specialties.

Empirical Finding 5: Panel members of the same medical specialty are significantly more likely to be linked than would be expected under chance.

Anecdotal experience also shows a relation between voting behavior and linkage patterns. If people who vote the same way also share linguistic attributes, then this suggests that their attention may be directed towards something that drives their decision outcome. This further suggests the possibility of agreement on a relatively small number of reasons for either approval or non-approval. On the other hand, the absence of links between members who vote the same way suggests that there may be a high diversity of reasons for why individuals vote a certain way, combined with attempts by some panel members to convince others who might disagree. In a similar manner to how we define specialty cohesion, we define *vote cohesion* as the proportion of edges in a graph that connect two panel members who vote the same way. *Vote cohesion percentile* is the proportion of random graphs, out of 1000 samples, that have lower vote cohesion than a graph representing a given meeting. There are 11 meetings in which there is a voting minority that has at least two people in it. These are used to generate a second meeting-specific distribution found in (in red) in Figure 47. This is contrasted against the vote cohesion percentile distribution for 1000 random graphs – a uniform distribution.

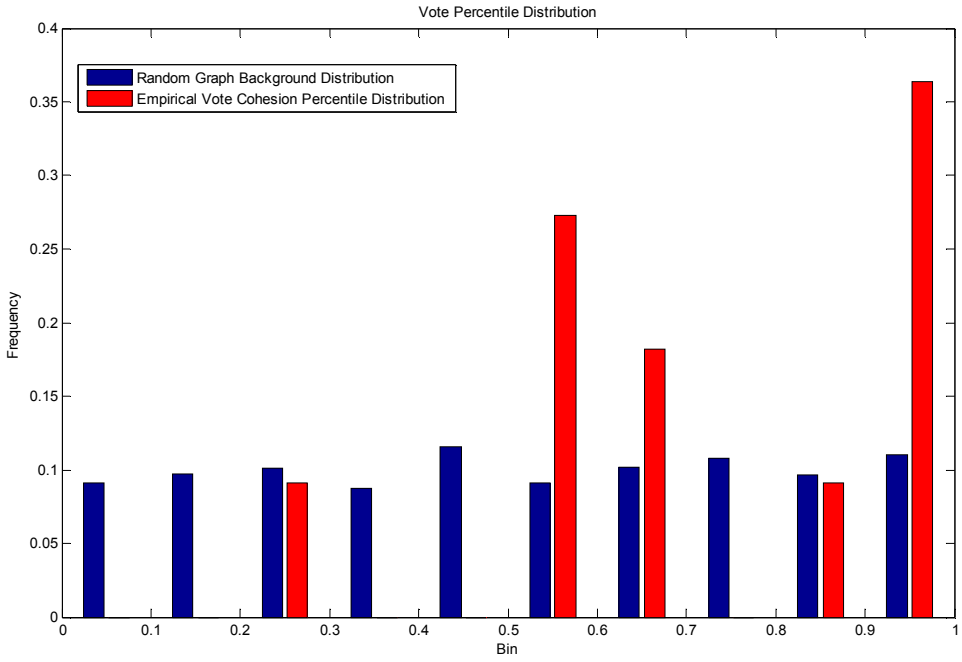


Figure 47: Histogram of Vote Cohesion Percentiles for the 11 meetings with a minority of size 2 or greater. The empirical vote cohesion percentile distribution's cumulative distribution function is significantly less than that of the background distribution (one-sided Kolmogorov-Smirnov test; p=0.015) indicating that the empirical distribution has more probability density concentrated near unity and away from zero.

We may see, by inspection, that the empirical vote cohesion percentile distribution has a right skew – i.e., probability mass is concentrated near 1 and away from 0. This suggests that people who vote alike are more likely to group together than we might expect under conditions of chance. A Kolmogorov-

Smirnov test for equality of distributions finds that the empirical cumulative distribution function (CDF) is significantly less than the uniform background CDF (p=0.015), shown in Figure 48. These results provide support for the notion that panel members who vote similarly tend to be linked.

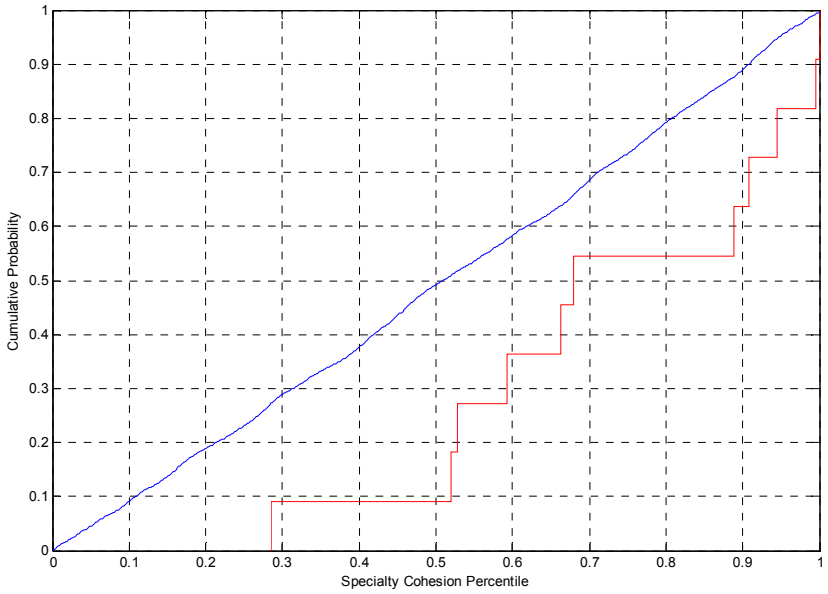


Figure 48: Cumulative Distribution Plot of Specialty Cohesion Percentiles for the 11 meetings with a minority with two or more voting members. The empirical vote cohesion percentile distribution's cumulative distribution function is significantly less than that of the background distribution (one-sided Kolmogorov-Smirnov test; p=0.015) indicating that the empirical distribution has more probability density concentrated near unity and away from zero.

Empirical Finding 6: Panel members who vote the same way are significantly more likely to be linked than would be expected under chance.

A scatter plot of specialty cohesion percentile vs. vote cohesion percentile for the 11 meetings analyzed shows that the two quantities are correlated (Spearman rho = 0.79, p=0.0061). This is a relatively tight correlation, suggesting strongly that specialty cohesion and voting cohesion increase together. In other words, meetings in which individuals' language links them by specialty are also meetings in which they are linked by vote. Of the 11 meetings observed, five have particularly high specialty cohesion and high vote cohesion, suggesting that these factors are dominant in these particular meetings.

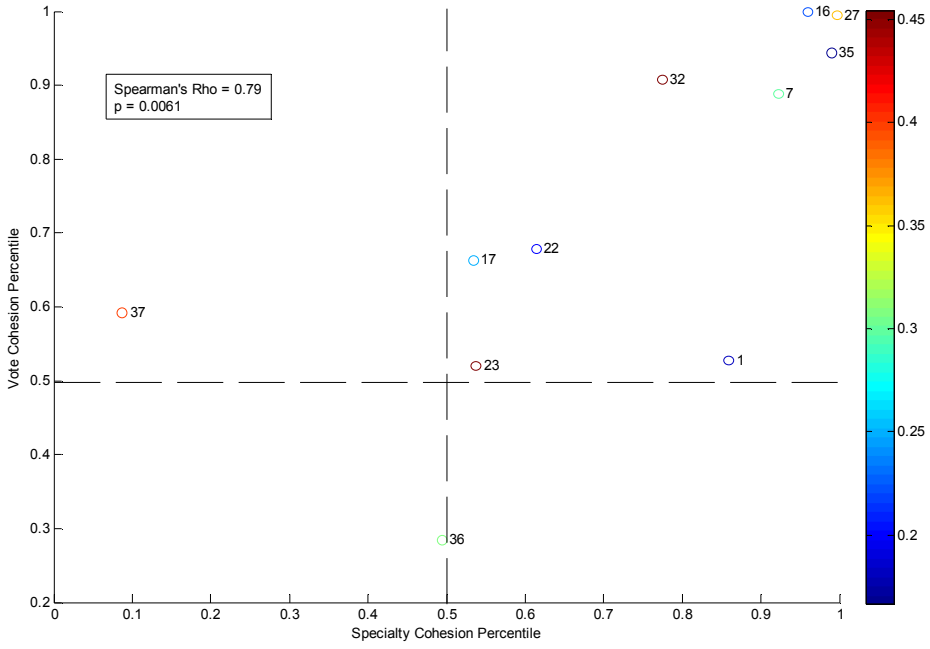


Figure 49: Scatter plot of Vote Cohesion percentile vs. Specialty Cohesion percentile for 11 meetings in

which there was a minority of two or more. Vote and specialty cohesion percentiles are positively associated (Spearman Rho =0.79; p=0.0061). Each datapoint is labeled by its corresponding meeting ID, as catalogued in Appendix 3. Datapoints are also color-coded by the proportional size of the minority in each meeting, suggesting that this effect holds independent of proportional minority size.

Empirical Finding 7: Vote cohesion percentile and specialty cohesion percentile are significantly positively associated for the subset of 11 meetings with at least two members in the voting minority.

Directed Graph Results

The Effects of Panel Member Speaking Order

Chapter 4 outlined a methodology for creating directed graphs by taking advantage of temporal ordering among topics. Influential panel members, who initiate topics that others follow, are more likely to be near the "top" of the graph (i.e., a low indegree) whereas panel members who are not followed tend to be near the "bottom" (low outdegree). This perhaps reflects a tendency for members of the voting minority to speak later in the meeting, compared to members of the voting majority. Figure 50 shows the difference between the median speaking order locations of voting majority and voting minority members. Voting minority members tend to speak later (p=0.0008) than do members of the voting majority.

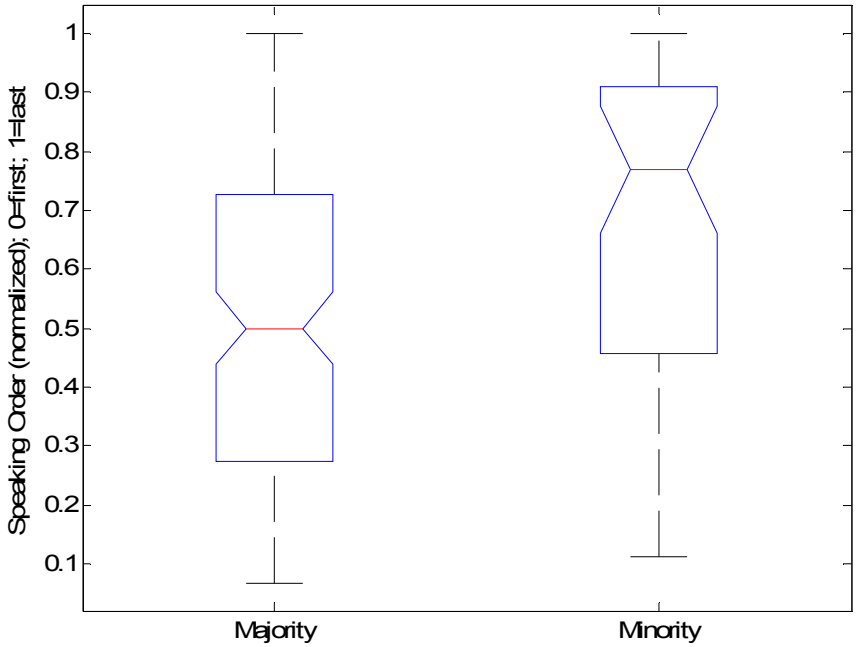


Figure 50: Kruskal-Wallis non-parametric ANOVA finds a significant difference between the median speaking order rank of voting majority and voting minority voting members in the 17 meetings in which there was a voting minority (abstentions were not included); p=0.0008. Voting minority members speak later than majority members do.

When meetings with a voting minority of only one member are excluded, we also find a significant difference between the median speaking order locations of members of the majority and the minority (p=0.011). A boxplot of this result is shown in Figure 51.

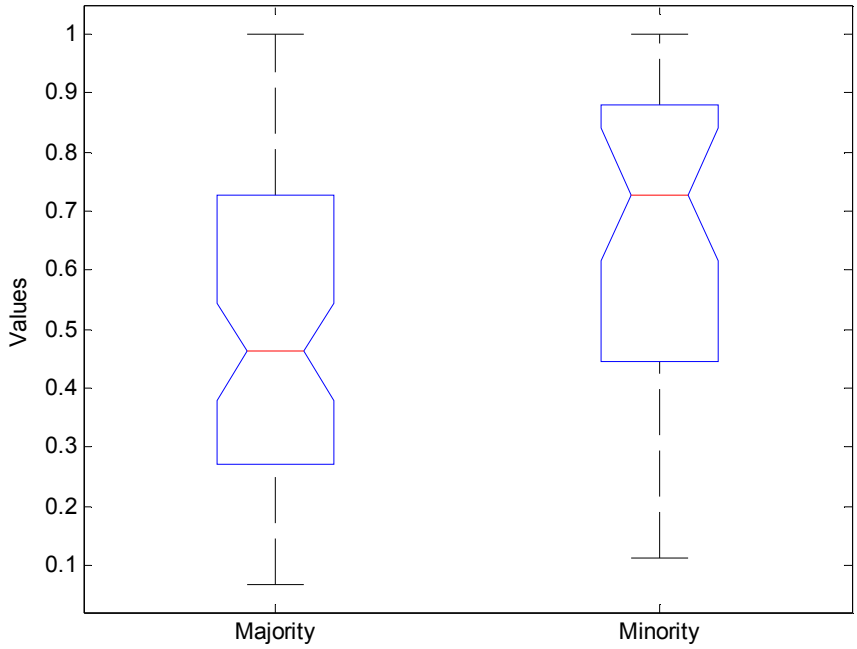


Figure 51: Kruskal-Wallis non-parametric ANOVA finds a significant difference between the median speaking order rank of voting majority and voting minority voting members in the 11 meetings in which there was a voting minority with two or more voting members (abstentions were not included); p=0.011. Voting minority members speak later than voting majority members do.

Empirical Finding 8: Members of the voting minority tend to speak later than do members of the voting majority.

Members of the voting minority tend to have a lower graph outdegree than do members of the voting majority (p=0.045), shown in Figure 52.

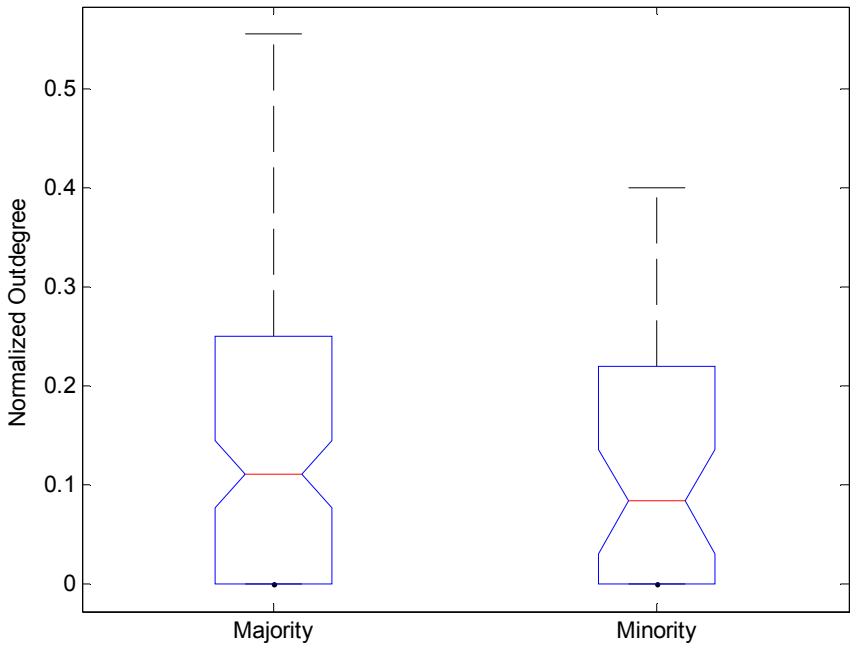


Figure 52: Kruskal-Wallis non-parametric ANOVA finds a significant difference between the outdegree of voting majority and voting minority panel members in the 17 meetings in which there was a majority (abstentions were not included); p=0.045. There is no observable effect for indegree (p=0.67) or undirected degree (p=0.37).

Examining the subset of 11 meetings in which there was a voting minority of size two or larger, we find that this effect also holds, but is only marginally statistically significant (p=0.058), shown in Figure 53.

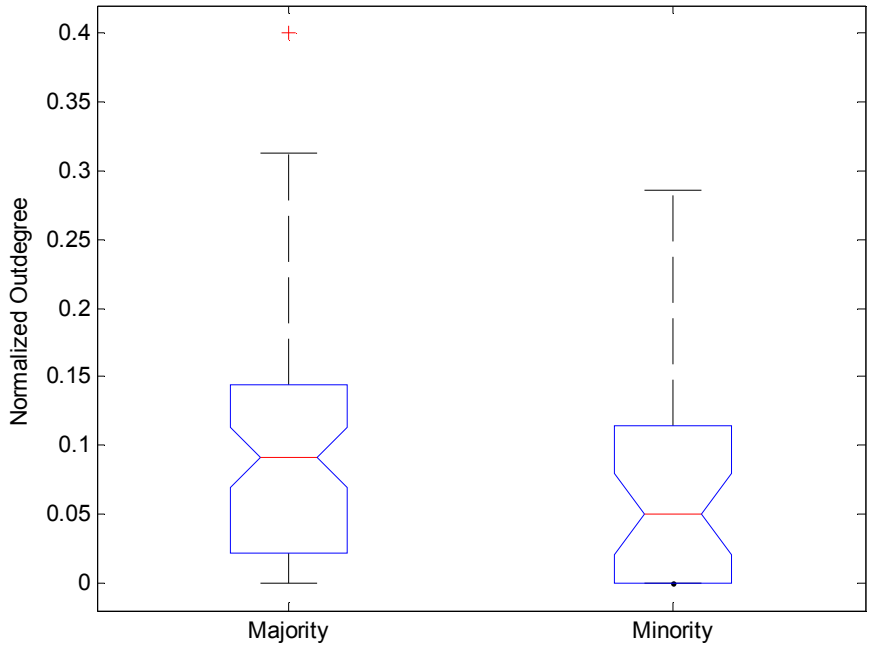


Figure 53: Kruskal-Wallis non-parametric ANOVA finds a significant difference between the outdegree of voting majority and voting minority panel voting members in the 11 meetings in which there was a majority of size two or larger (abstentions were not included); p=0.058.

Empirical Finding 9: Members of the voting minority tend to have a lower graph outdegree than do members of the voting majority.

There is an association between outdegree and speaking order (Spearman rho=-0.35; $p=1.15 \times 10^{-6}$), and between indegree and speaking order (Spearman rho=0.45; $p=5.9 \times 10^{-11}$) for the 17 meetings with a voting minority, something that is to be expected given that directionality is chosen on the basis of topic-

ordering, which is in turn shaped by procedural constraints. This association also holds for the subset of 11 meetings with a minority of two or more (Outdegree rho = -0.27; p=0.0026); (Indegree rho = -0.48; $p=5.9 \times 10^{-8}$). Furthermore, an analysis of covariance shows no significant difference between the correlations between location in speaker order and meetings with a minority of one compared to meetings with a minority of two or more (p=0.49 for outdegree; p=0.34 for indegree).

Empirical Finding 10: Outdegree is negatively and significantly associated with location in the speaking order, and indegree is positively and significantly associated with location in the speaking order.

An ANOVA identifies speaking order as capturing the main effect in voting behavior, whereas the effect due to outdegree is not significant (see Table 11)..

Table 11: 2-way ANOVA Table showing effect of Outdegree and Speaking Order on vote (majority vs. minority) for those 17 meetings in which there is a minority. Although the ANOVA assumptions are not met, an effect of Speaking Order is still evident (cf. Lunney 1970). The absence of an effect due to outdegree suggests that the variance in speaking order accounts for the variance in voting behavior as well as in outdegree.

Variable	Sum of Squares	Degrees of Freedom	Mean Squares	F	p-value
Outdegree	0.076	1	0.076	0.043	0.51

(normalized)					
Speaking Order (normalized)	1.5	1	1.5	8.7	0.0036
Error	30	173	0.18		
Total	32	175			

This result also holds for the subset of 11 meetings with a voting minority of at least two voting members, as shown in Table 12.

Table 12: 2-way ANOVA Table showing effect of Outdegree and Speaking Order on vote (voting majority vs. voting minority) for those 11 meetings in which there is a minority with at least two members. Although the ANOVA assumptions are not met, an effect of Speaking Order is still evident. The absence of an effect due to Outdegree suggests that the variance in speaking order accounts for the variance in voting behavior as well as in outdegree.

Variable	Sum of Squares	Degrees of Freedom	Mean Squares	F	p-value
Outdegree (normalized)	0.082	1	0.082	0.28	0.60

Speaking Order (normalized)	1.44	1	1.44	4.91	0.029
Error	34.6	118	0.29		
Total	36.6	120			

Empirical Finding 11: Location in the speaking order seems to account for the variance in voting behavior that is associated with outdegree.

Recall that empirical finding 2 shows that vote and air-time are not associated. Instead, some of the above results might seem to indicate that the outcome of a meeting depends on a procedure that could equalize air-time, but that might have other effects on voting behavior. To further explore this idea, we examined the subset of meetings in which there was a minority of at least one person. These meetings were then further subdivided into meetings in which the device was approved (n=7) and meetings in which the device was not approved (n=10). In meetings in which the device was not approved, we found that members of the voting majority (i.e., those who voted against device approval) spoke significantly earlier than did members of the voting minority (p=0.0025; see Figure 54)

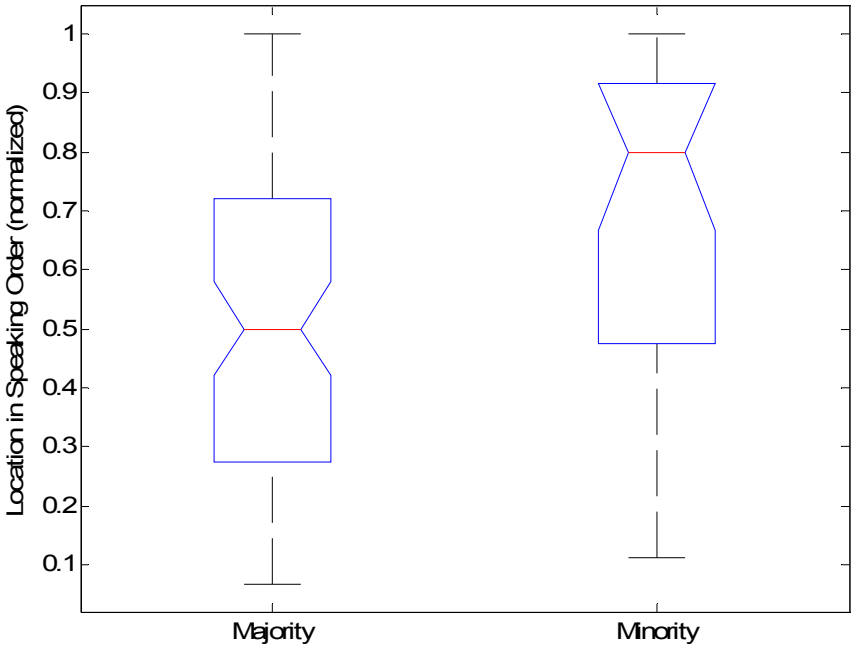


Figure 54: Members of the voting minority (in favor of device approval) speak significantly later than do members of the voting majority (against device approval) in the 10 meetings in which the panel voted not to approve the devices (p=0.0025)

In meetings in which the device was approved, there was no significant difference between members of the voting majority and voting minority (p=0.12), although there is a trend towards members of the voting minority speaking later (see Figure 55).

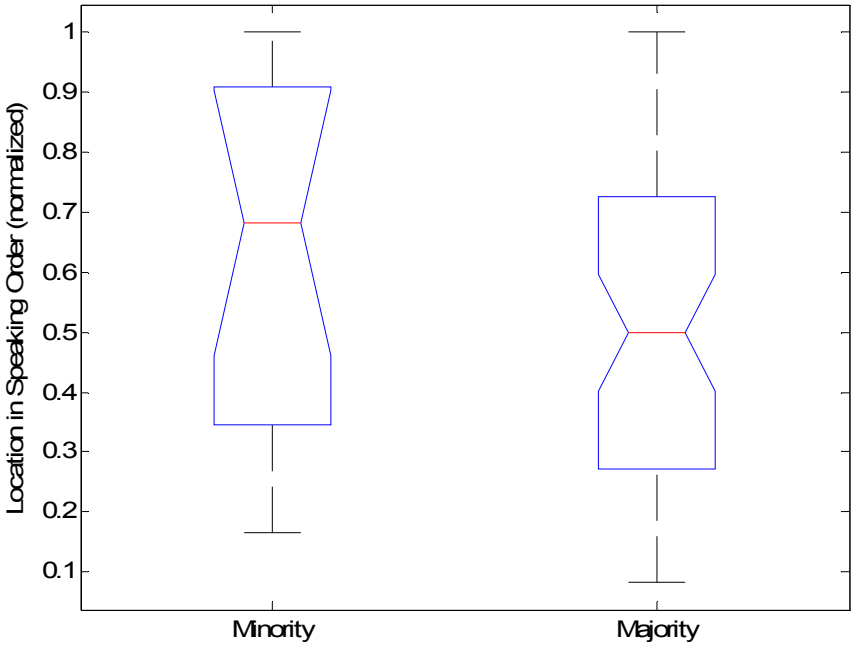


Figure 55: Members of the voting minority (against device approval) do not speak significantly later than do members of the voting majority (in favor of device approval) in the 7 meetings in which the panel voted not to approve the devices ($p=0.12$). By inspection, there is a non-significant trend for the voting minority to speak later than does the voting majority.

Empirical Finding 12: Members of the voting minority spoke significantly later in meetings in which the panel did not approve the devices than did members of the voting majority. This trend was not present in meetings in which the panel did approve the device.

An opposite trend was found when examining outdegree. In meetings in which the device was not approved (n=10), we found that there was no significant difference between the normalized outdegrees of members of the voting majority (i.e., no voters) and members of the voting minority (p=0.27; see Figure 56).

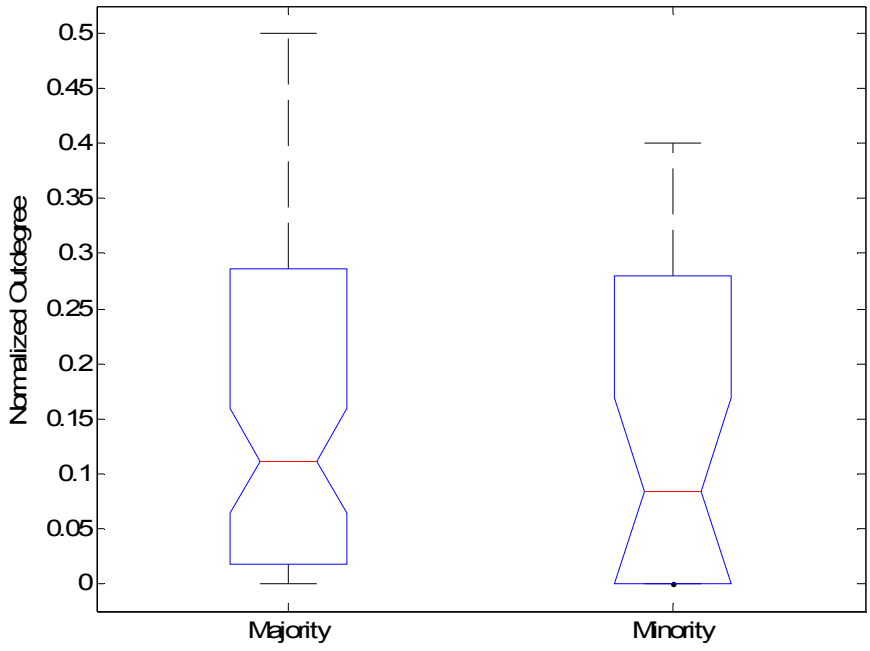


Figure 56: Members of the voting minority (in favor device approval) do not have significantly smaller outdegrees than do members of the voting majority (against device approval) in the 10 meetings in which the panel voted not to approve the devices (p=0.27).

In meetings in which the device was approved (n=7), there was a marginally significant difference between members of the voting majority and voting

minority (p=0.056), such that members of the voting majority had a higher normalized outdegree.

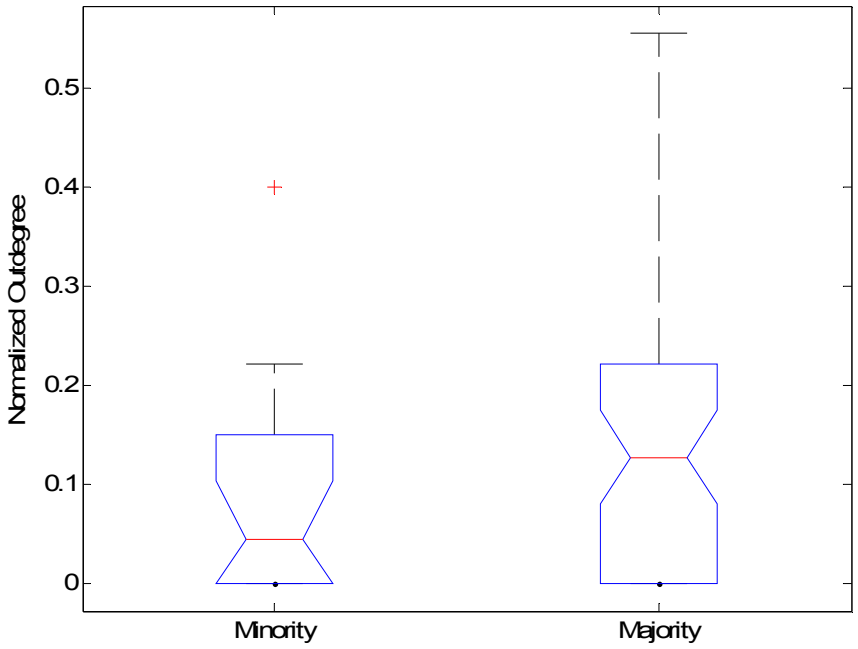


Figure 57: Members of the voting minority (against device approval) have marginally significantly smaller outdegrees than do members of the voting majority (in favor of device approval) in the 7 meetings in which the panel voted to approve the devices (p=0.056).

Empirical Finding 13: Members of the voting minority had a significantly smaller outdegree in meetings in which the device was approved. This trend was not present in meetings in which the panel did not approve the device.

Analysis shows that use of directed graphs can help identify which voting members are likely to be part of the voting minority. This is accomplished by analyzing "graph sinks", i.e., nodes with zero outdegree and nonzero indegree. Graph sinks are more likely to be members of the voting minority than are other nodes (see Table 13).

Table 13: Analysis of the 17 meetings with a voting minority indicates that members of the minority are more likely to be graph sinks than are members of the majority ($\chi^2 = 4.92$; dof=1; p=0.026).

	Sink	Non-Sink	TOTAL
Voting Minority	13	30	43
Voting Majority	20	113	133
TOTAL	33	143	176

This result also holds for the subset of meetings in which there is a majority including at least two voting members, as shown in Table 14.

Table 14: Analysis of the 11 meetings with a voting minority including at least two members indicates that members of the voting minority are more likely to be graph sinks than are members of the voting majority ($\chi^2 = 4.66$; dof=1; p=0.031).

	Sink	Non-Sink	TOTAL
Voting Minority	11	26	37
Voting Majority	10	67	77
TOTAL	21	93	114

Empirical Finding 14: Members of the voting minority are more likely to be graph sinks than are members of the voting majority.

These results are understandable in light of the speaking-order effect on FDA panels identified above. In particular, we find that panel members who speak last are more likely to be in the voting minority than are panel members who don't speak last (see Table 15).

Table 15: Analysis of the 17 meetings with a voting minority shows that members of the voting minority are more likely to the last speaker than are members of the voting majority (χ^2 =5.22; dof=1; p=0.022)

	Last Speaker	Other	TOTAL
Voting Minority	8	35	43
Voting Majority	9	124	133
TOTAL	17	159	176

Although a voting minority member is almost twice as likely to be the last speaker as is a voting majority member in the subset of 11 meetings in which there is a voting majority including at least two panel members, this result is not statistically significant, as shown in Table 16. On the other hand, singleton voting minority members are significantly more likely to be the last speaker than are members of the much larger voting majority, as shown in Table 17.

Table 16: Analysis of the 11 meetings with a voting minority of size two or more shows that members of this voting minority are not more likely to the last speaker than are members of the voting majority (χ^2 =0.94; dof=1; p=0.33)

	Last Speaker	Other	TOTAL
Voting Minority	5	32	37
Voting Majority	6	71	77
TOTAL	11	103	114

Table 17: Analysis of the 6 meetings with a voting minority of size one only shows that members of the voting minority are more likely to the last speaker than are members of the voting majority (χ^2 =12.36; dof=1; p=0.00044). Of the three voting minority members who are the last

speaker, two are graph sinks and one is a graph isolate (outdegree and indegree are both 0).

	Last Speaker	Other	TOTAL
Voting Minority	3	3	6
Voting Majority	3	53	56
TOTAL	6	56	62

Empirical Finding 15: Members of the voting minority are more likely to be the last speaker to ask questions of the sponsor and FDA, than are members of the voting majority, especially for meetings in which there is a singleton voting minority.

We therefore have two competing heuristics that might be used to evaluate whether a given voter is likely to be in the minority. This is a binary classification task, whose efficacy we can measure using the "F-score", a commonly used metric in the information retrieval literature. The F-score is defined as the harmonic mean of precision and recall, where:

$$\text{Precision} = \frac{(TruePositives)}{(TruePositives + FalsePositives)} \quad \text{and:}$$

$$\text{Recall} = \frac{(TruePositives)}{(TruePositives + FalseNegatives)}$$

Therefore:

$$F = 2 \times \frac{(\mathrm{Pr}\,ecision \times \mathrm{Re}\,call)}{(\mathrm{Pr}\,ecision + \mathrm{Re}\,call)}$$

Table 18 shows precision, recall and the F-score for the two conditions examined.

Table 18: Table of Precision, Recall, and F-Score for the data shown above. The graph sink method has a consistently higher precision and F-score, and is lower on recall only in the case of 17 meetings.

	17 Meetings with a minority of size 1 or larger		11 Meetings with a minority of size 2 or larger	
	Graph Sinks	Last Speaker	Graph Sinks	Last Speaker
Precision	0.30	0.19	0.30	0.14
Recall	0.40	0.47	0.52	0.45
F-score	0.34	0.27	0.36	0.21

We note that precision and F-score are both higher for the graph sink heuristic across both conditions. Recall is higher for the last speaker condition only when the six meetings with a voting minority of size one are included.

Empirical Finding 16: Using F-score as an evaluation criterion, the graph sink heuristic provides a superior classification of minority members when compared to the last speaker heuristic.

The Effects of Lead Reviewers

Speaking order is an important variable associated with voting behavior. Thus we would also like to examine those at the start of the speaking order – namely, the lead reviewers. Lead-reviewers are panel members designated by FDA to review a given device in more depth prior to the panel proceedings. They always speak first or immediately after another lead reviewer. A panel meeting may have as many as two lead reviewers. Although lead reviewers speak more than do other voting members across the set all subsets of meetings[9], lead reviewers are not significantly more likely to be in the minority when compared to other voting members[10]. Although lead reviewers have a significantly or marginally-significantly larger outdegree[11] and a significantly smaller indegree[12] than do other panel members, lead reviewers in the minority do not have a significantly different outdegree or indegree from lead reviewers in the majority[13].

Empirical Finding 17: Although lead reviewers have a significantly higher air-time and outdegree, and a significantly lower indegree than other panel members, their overall voting behavior is not significantly different.

There is at least one lead reviewer in the majority for all but one meeting for which lead reviewers were assigned (n=35). The one outlier was a meeting in which the lead reviewer was a specialty isolate – i.e., the only surgeon on a committee largely composed of cardiologists. This individual was also a graph

[9] (n=37;. p=0.031 by a Kruskal-Wallis analysis); (n=17; p=0.0004 by a Kruskal-Wallis analysis); (n=11; p=0.0115 by a Kruskal-Wallis analysis)

[10] (n=37;.χ^2=0.27; dof=1; p=0.87); (n=17;.χ^2=0.0029; dof=1; p=0.95); (n=11;.χ^2=0.15; dof=1; p=0.70)

[11] (n=37;. p=2.1 x 10^{-5} by a Kruskal-Wallis analysis); (n=17; p=0.011 by a Kruskal-Wallis analysis); (n=11; p=0.079 by a Kruskal-Wallis analysis)

[12] (n=37;. p=1.31 x 10^{-5} by a Kruskal-Wallis analysis); (n=17; p=0.0017 by a Kruskal-Wallis analysis); (n=11; p=0.0026 by a Kruskal-Wallis analysis)

[13] (n=37;. p=0.32 by a Kruskal-Wallis analysis); (n=17; p=0.22 by a Kruskal-Wallis analysis); (n=11; p=0.42 by a Kruskal-Wallis analysis)

isolate (i.e. indegree and outdegree = 0). Furthermore, for meetings where there is at least one lead reviewer in the voting minority, the voting minority tends to be proportionally larger (p=0.0006), shown in Figure 58.

Figure 58: Kruskal-Wallis non-parametric ANOVA finds a significant difference between proportional voting minority size in the 35 meetings in which there was a voting minority and at least one lead reviewer in the voting minority (p=0.0006). Similar results are obtained when focusing on the subset of 17 meetings with a voting minority (p=0.027). There is insufficient data to obtain a similar result for the subset of 11 meetings with a voting minority of 2 or more (p=0.33), although the

direction of the trend remains the same.

Furthermore, with only one exception, in cases in which there was at least one lead reviewer in the voting minority, there was a second lead reviewer in the voting majority. The exception was the meeting held on July 9, 2001 (Meeting ID: 16) in which the lead reviewer was a surgeon on a committee comprised largely of cardiologists. Therefore, the proportional size of the voting minority is also larger when lead reviewers disagree[14].

Empirical Finding 18: The proportional size of the voting minority is larger when lead reviewers do not vote with the majority within a given meeting, and when there is disagreement among lead reviewers.

Furthermore, we find that meeting length is negatively correlated with the proportion of lead reviewers in the voting majority (Spearman Rho = -0.37, p=0.027).

Empirical Finding 19: Meetings are longer when more lead reviewers are in the voting minority.

Although, in general, members of the voting minority are more likely to be graph sinks, we find that, as meetings get longer, the maximum outdegree of a member of the voting minority increases – i.e., a voting minority member is more likely to reach the "top" of the graph (Spearman rho = 0.50; p=0.04). Figure 59 shows the relation between the maximum outdegree of a member of the voting minority and meeting length. This is consistent with the anecdotal time dependence seen in Chapter 3 wherein graphs of later of meeting subsections showed increasing connectivity across voting blocs.

[14] (p=0.0015; n=35), (p=0.029; n=16), (p=0.20; n=10, ns)

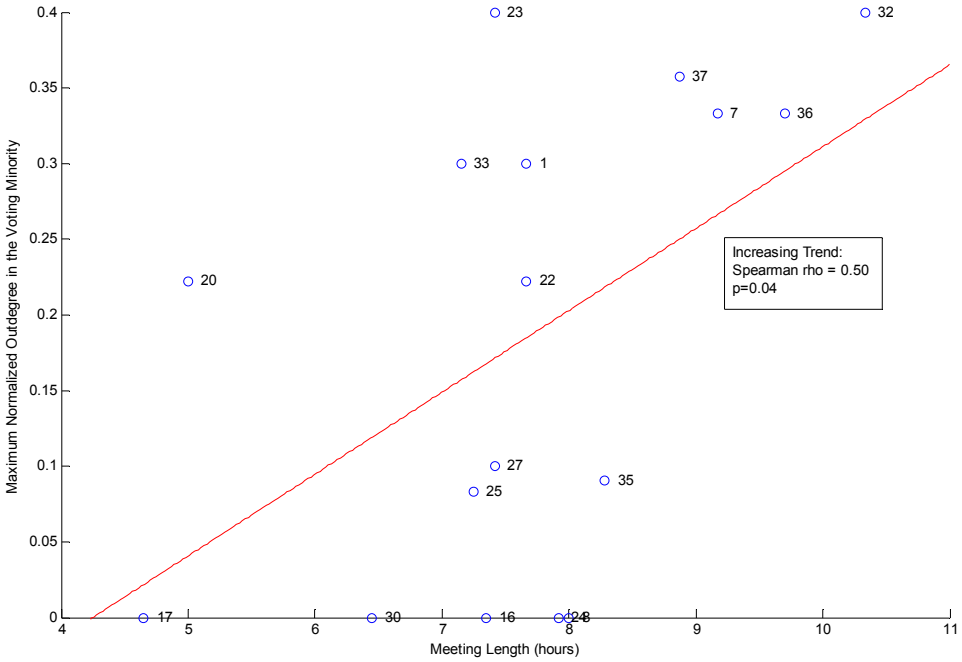


Figure 59: Maximum normalized outdegree is significantly associated with meeting length (Spearman rho=0.50; p=0.04). Datapoints are labeled by the meeting ID assigned in Appendix 3. There is no significant association between location of first minority member in the speaking order and meeting length (p=0.50).

Empirical Finding 20: Meeting length is significantly positively associated with the maximum normalized outdegree among voting minority members, but not with maximum location in the speaking order.

Under such conditions, voting minorities also become larger – indeed, the maximum outdegree of a voting minority member is strongly associated with the

proportion of voting members in the minority (Spearman rho = 0.62, p=0.0082; see Figure 60)

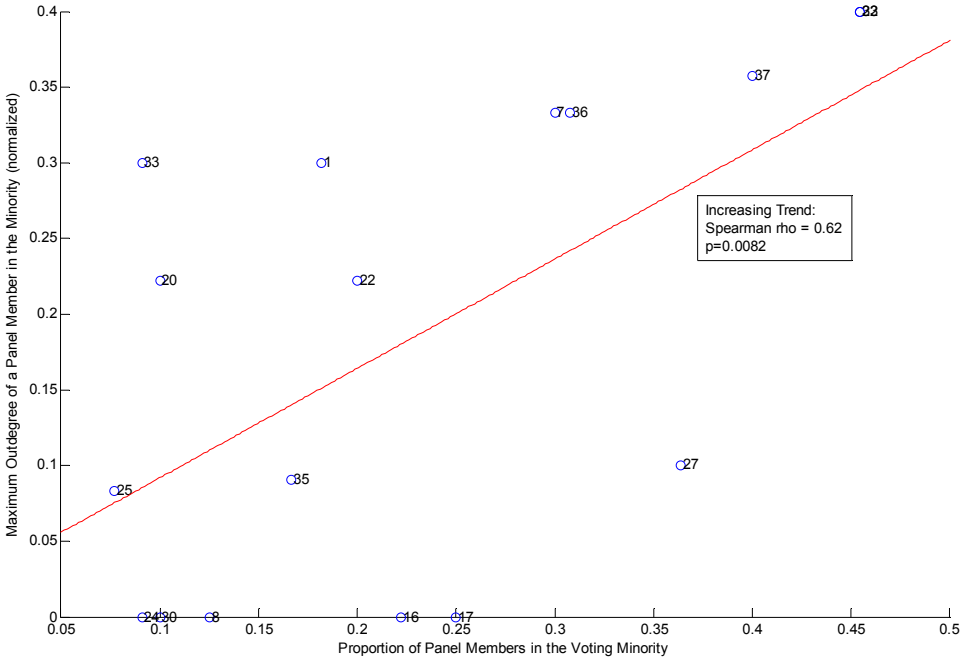


Figure 60: Maximum normalized outdegree is significantly associated with voting minority proportional size (Spearman rho=0.62; p=0.0082) for the 17 meetings in which there is a minority. Datapoints are labeled by the meeting ID assigned in Appendix 3.

Empirical Finding 21: Maximum normalized outdegree is significantly associated with proportional voting minority size.

This is generally consistent with the observation that as meeting length increases, so does the size of the voting minority (see Figure 61).

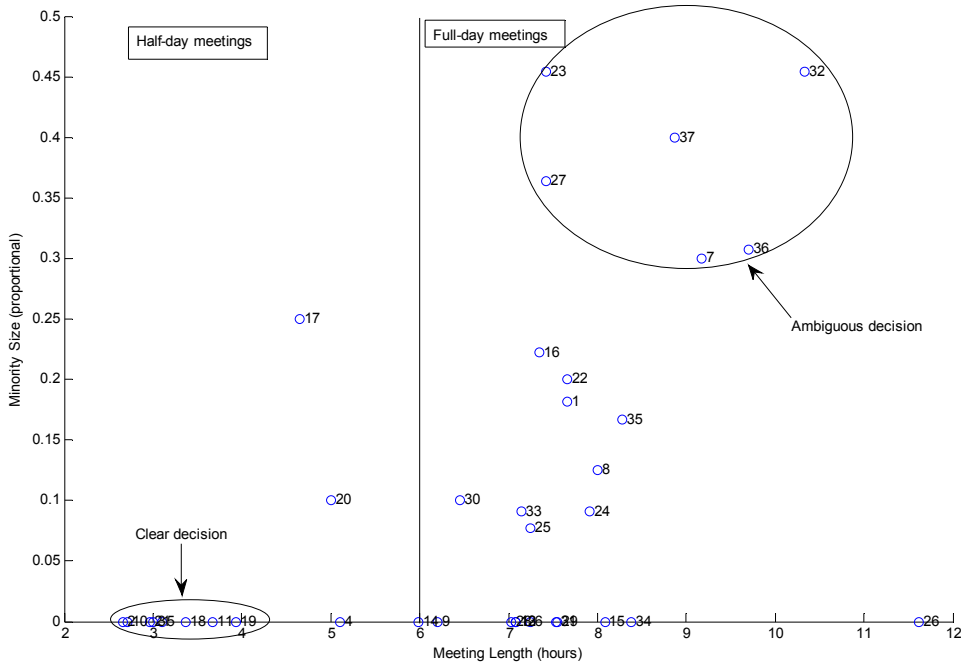


Figure 61: Plot of Meeting Length vs. voting minority proportional size. Meeting length is significantly positively associated with voting minority proportional size (Spearman Rho = 0.53; $p=7.1 \times 10^{-4}$). Decisions that are likely to have been clear or ambiguous are labeled.

Empirical Finding 22: Meeting length is significantly positively associated with proportional voting minority size.

One possible interpretation of this result is that longer meetings are associated with difficult decisions, perhaps due to complex devices or procedures, poor data quality or other sources of ambiguity about the device. Longer meetings typically involve more committee deliberation, which is more likely to be necessary when there is no consensus on how best to interpret the available data. In these cases,

minority voters tend to be more randomly distributed in the graphs and procedural effects seem minimal.

Chair effects

The results shown above do not include the impact of the committee chair in the analysis. Using the directed graphs developed in the previous chapter, we can determine the impact of the committee chair on the meeting by examining his/her role in facilitating communication. Figure 62 shows a directed graph from the meeting held on June 23, 2005.

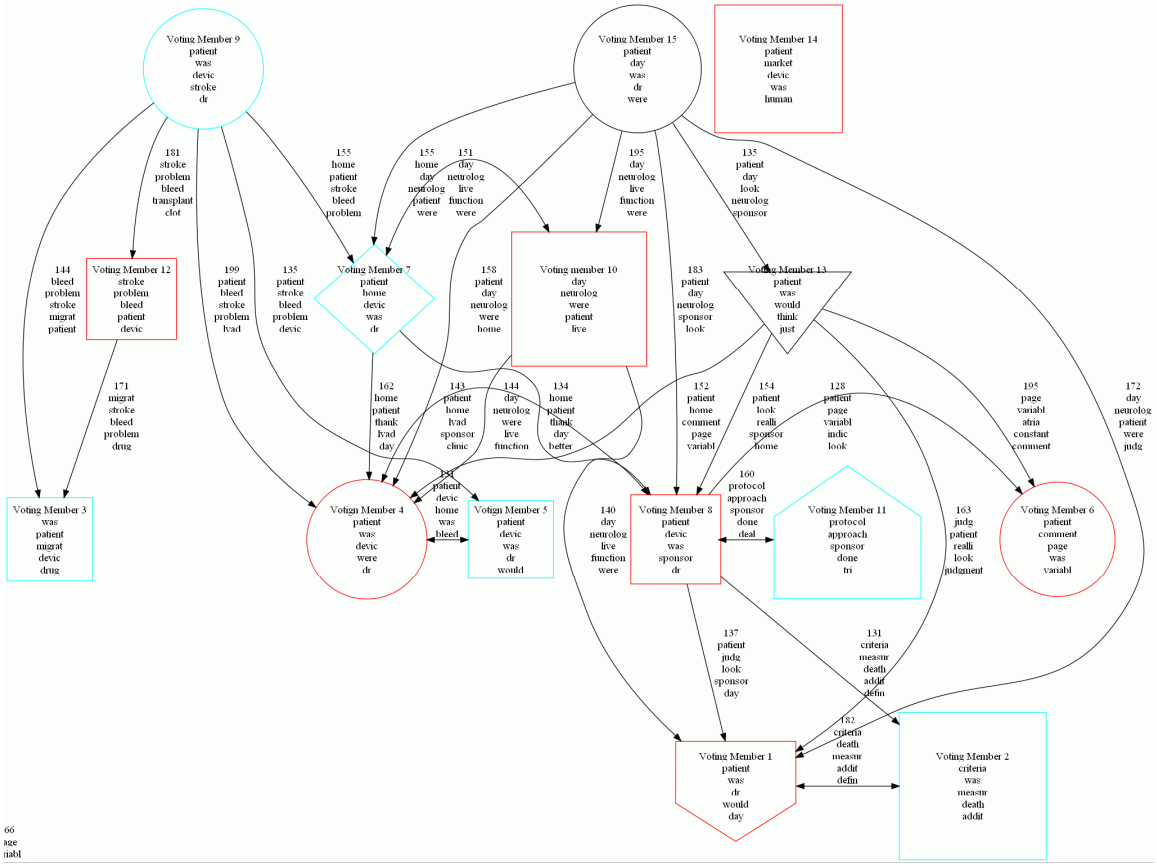


Figure 62: Directed Graph representation of meeting held on June 23, 2005. Luo's hierarchy metric = 0.35.

We may quantify the impact that the committee chair has upon the meeting by determining, for each graph, the proportion of edges which are part of a cycle. This is a metric of the hierarchy in the graph (Luo et al., 2009). We display this metric for the graph without the chair (e.g., Figure 62) and with the chair (e.g., Figure 63).

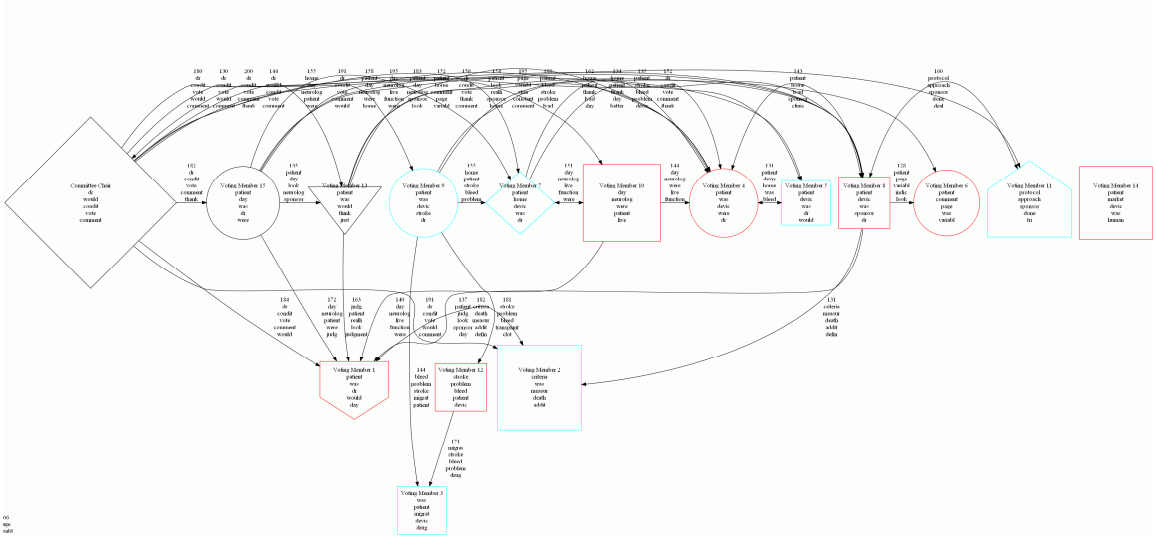


Figure 63: Directed Graph representation of meeting held on June 23, 2005, with the committee chair included. Luo's hierarchy metric = 0.78.

The difference in this metric between graphs with and without the chair therefore quantifies the impact of the chair on the meeting. For the meeting held on June 23, 2005, this value is 0.78-0.35 = 0.43. This suggests that the chair is significantly changing the topology of the meeting structure – in particular, it seems that the chair is connecting members at the "bottom" of the graph to those at the "top".

Other meetings display different behavior by the chair. Consider the meeting held on October 27, 1998 (in Figure 64 and Figure 65).

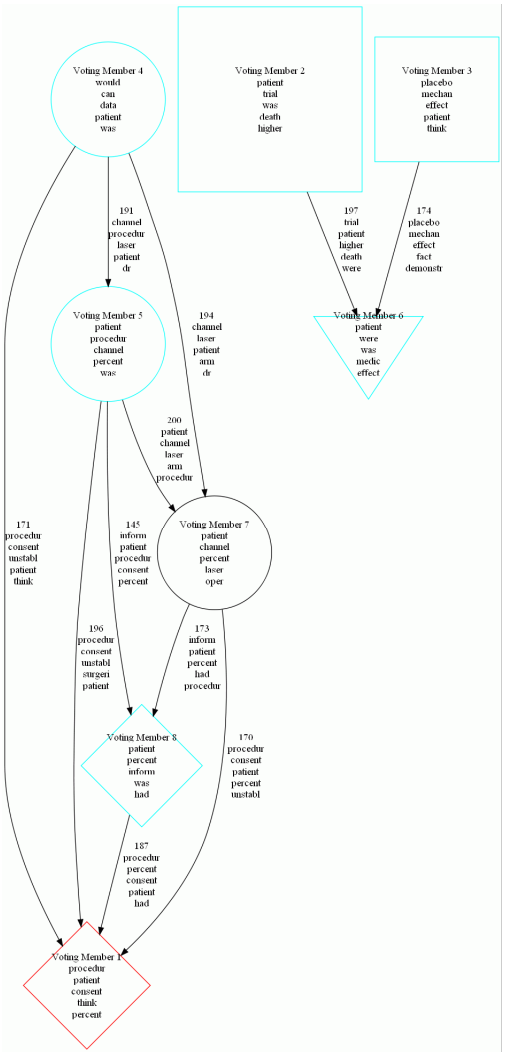


Figure 64: Directed Graph representation of meeting held on October 27, 1998. Luo's hierarchy metric = 0.

In this meeting, the committee chair served to connect the two disparate clusters on the panel. Nevertheless, the chair is not creating any new cycles on the graph. This is reflected in the fact that the hierarchy metric for both of these meetings is equal to 0.

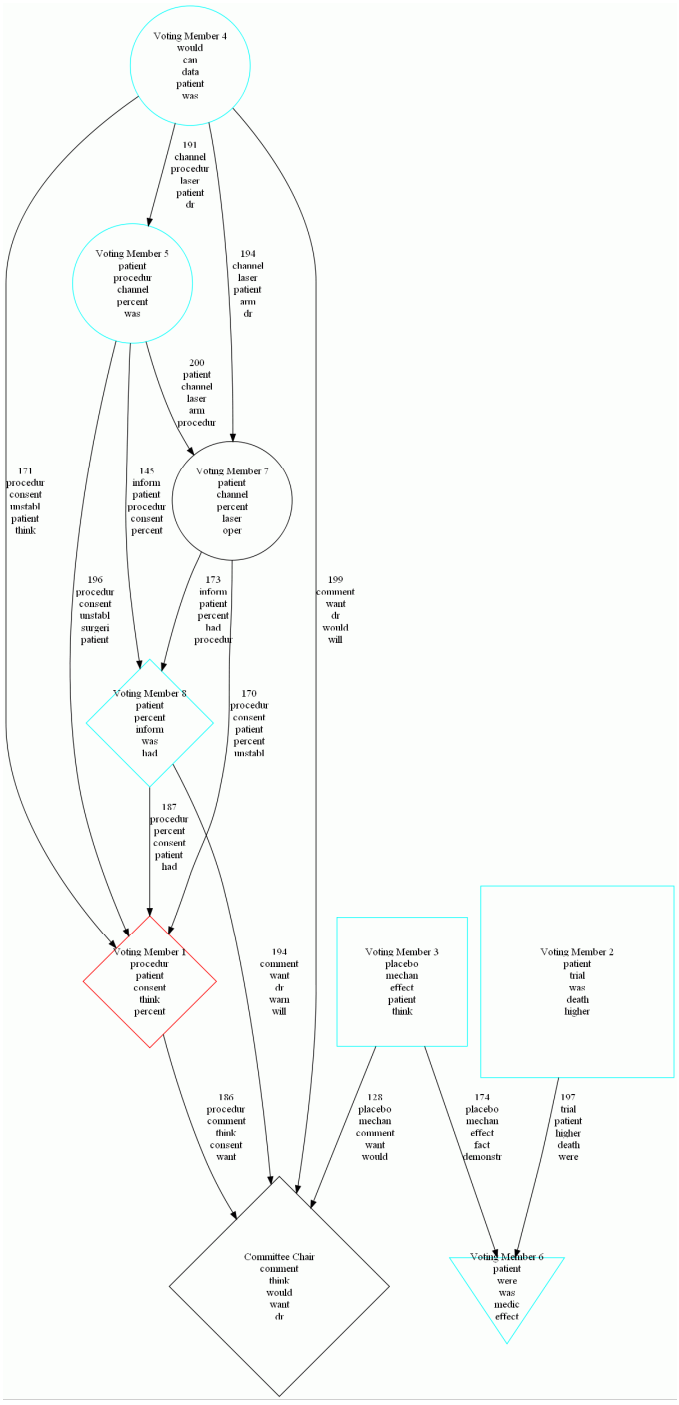


Figure 65: Directed Graph representation of meeting held on October 27, 1998, with the committee

chair included. Luo's hierarchy metric = 0.

Given that both meetings have a hierarchy of 0, the difference between them is also 0. In general, we can examine the difference in hierarchy for a given meeting. A histogram of these is shown in Figure 66. This histogram is bimodal.

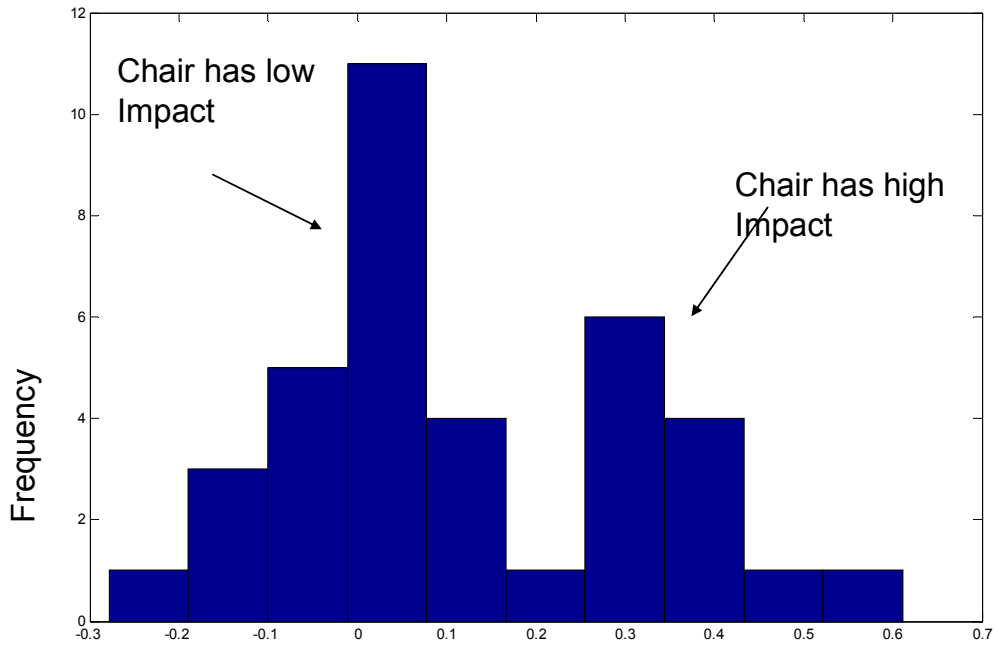


Figure 66: Distribution of chair impacts for the set of 37 meetings analyzed. This distribution shows a bimodal structure.

Empirical Finding 23: Inclusion of the Committee Chair in directed graphs leads to a bimodal distribution of the extent to which the chair changes the structure of the graph. These two modes may correspond to

different sorts of behavior by the Chair in his/her interactions with panel members during the meeting.

This bimodal structure is particularly pronounced when we focus on the subset of meetings in which there is a voting minority. Among these meetings, the bimodality seems to be associated with meeting date (p=0.005). This effect is shown in Figure 67.

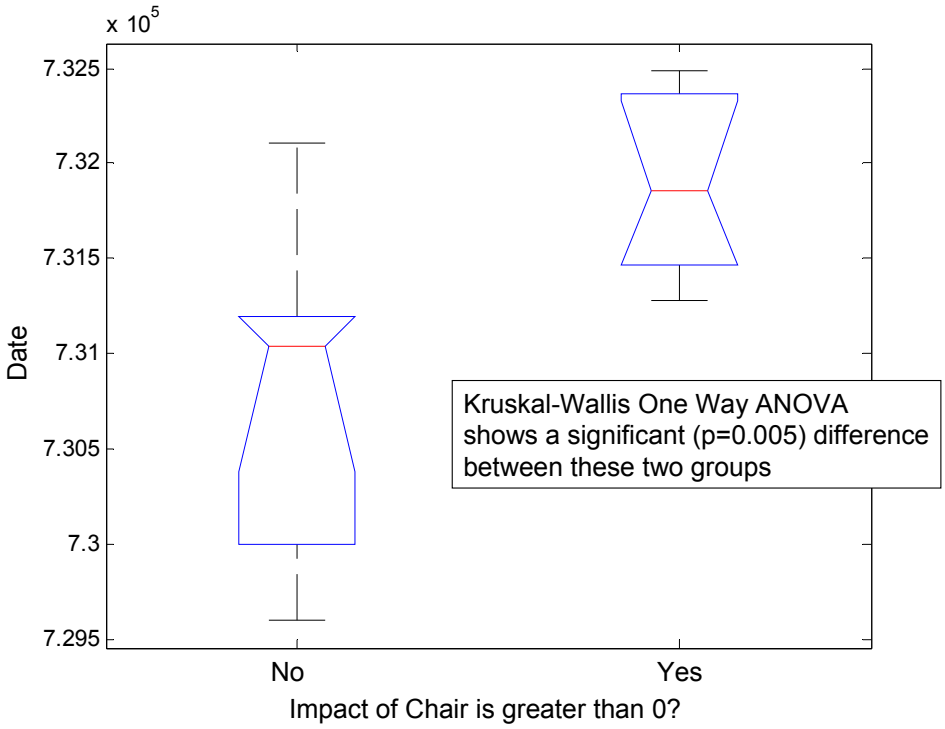


Figure 67: The impact of the committee chair seems to be associated with meeting date. The vertical axis represents the number of days since January 1st, 1900.

Empirical Finding 24: Committee chair impact is significantly positively associated with meeting date for meetings in which there is a voting minority.

A closer analysis of the meetings associated with this distribution is instructive (see Figure 68). We see that the impact of the chair seems to be increase markedly around March 2002.

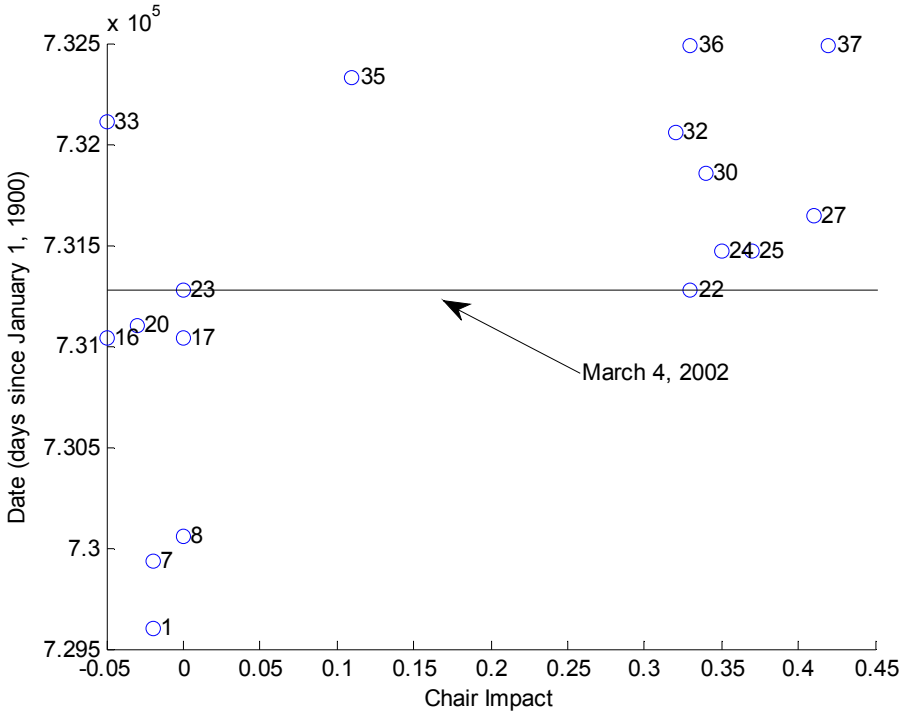


Figure 68: Impact of chair vs. meeting date for each of the 17 meetings in which there was a voting minority. Note that after March 4, 2002, chair impact seems to increase for most meetings. Each meeting is labeled by its corresponding ID.

Using this date as a cutoff, we find that meetings prior to March 4, 2002 are significantly shorter than meetings after March 4, 2002 ($p=0.0004$, by a Kruskal-Wallis test). This is largely because half-day meetings were no longer held after this date (see Figure 69).

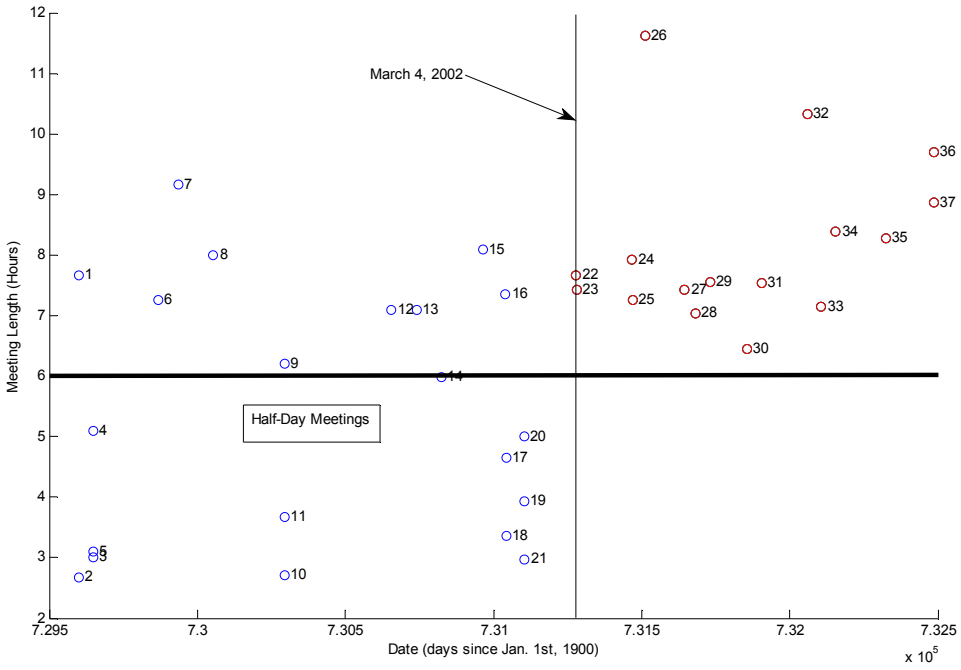


Figure 69: Half-day meetings were not held after March 4, 2002. These later meetings are marked in red, whereas earlier meetings are in blue. Each meeting is labeled by its corresponding ID.

Concerns about bias on FDA Panels

(Sherman 2004) identifies two potential sources of "bias" on FDA Panels. These may be broadly construed as financial and intellectual. Up until now, we have been largely studying intellectual factors, e.g., those related with medical specialty. A financial conflict of interest arises when a panel member in some way receives funding from either the device sponsor or one of its competitors. We would expect conflicts of interest to arise when panel members who possess the appropriate expertise yet lack a financial conflict are not available (McComas, Tuite & Sherman 2005). Indeed, we find that panel members with conflicts of interest have a higher h-index and therefore, or academic expertise, than do panel members without a conflict ($p=0.002$, by a Kruskal-Wallis test).

Empirical Finding 25: Panel members with conflicts of interest tend to have a significantly higher h-index than do panel members without a conflict.

Concerns regarding the effects that panel members with conflicts of interest might have on panel operations have been raised frequently with regards to FDA panels, particularly in the news media. A study by Lurie et al. (2006) found no significant relation between committee voting outcomes and conflict of interest on a subset of panels in the Center for Drug Evaluation and Research (CDER), and only a small relation between individual voting outcome and conflict of interest. One limitation of Lurie's work is its inability to account for influence patterns – in particular, a given panel member may influence the vote of another through direction of attention. The method presented in chapter 3 and applied in this chapter allows an analysis of this potential effect. If an individual with a conflict of interest is influencing other panel members, then that individual would have a relatively high outdegree. We find that this is not the case ($p=0.66$ using a Kruskal-Wallis test)

Empirical Finding 26: Panel members with conflicts of interest do not have higher outdegrees than panel members without conflicts of interest.

A possible concern is that when panel members with conflicts of interest are in influential positions (e.g., they have a high outdegree), the panel will follow them. We find that there is no significant difference between the outdegrees of panel members with conflicts who are in the voting majority and the outdegrees of panel members with conflicts who are in the voting minority ($p=0.38$).

Empirical Finding 27: Panel members with conflicts of interest, who are in the voting majority do not have higher outdegrees than panel members with conflicts of interest who are in the voting minority.

We find that, across our sample of 37 meetings, members with conflicts of interest are in the voting majority in 45 times out of 49 total conflicts (i.e, 92% of the time). After January 2002, panel members were required to report not only the presence or absence of a conflict of interest, but also its direction (i.e., with the sponsor or one of the sponsor's competitors). In the 15 meetings since the beginning of 2002, there were 34 total conflicts of interests, of which 30 (88%) were in the voting majority. This finding is offset by the fact that of those 34 conflicts of interest, 13 (38%) voted against their conflict of interest. There were 14 meetings in which there was at least one panel member with a reported direction of conflict of interest. Of these, there were four meetings in which there was only one member with a conflict. In each of these four meetings, this member either voted against the conflict reported or was not allowed to vote at all. In each case, the panel's voting outcome went against the reported conflict. There were five meetings in which there were multiple conflicts of interest that went in the same direction. In all but one of these meetings, the panel voted unanimously in favor of the direction consistent with the conflict. Finally, there were five meetings in which the conflicts of interest were "balanced" – i.e.,

members representing both directions were present. In these meetings, there were a total of 15 panel members with conflicts of interest. Seven of these voted in the direction of their conflict, and eight voted against the direction of their conflict. This evidence suggests that conflicts of interest may be minimized when there is only one member on the panel with a reported conflict or when there are opposing reported conflicts on the panel. When there are multiple panel members with consistent conflicts of interest, unanimous support for those conflicts might result. More data is required to rigorously test this finding.

This chapter presented 27 empirical findings derived from an application of statistical analysis and the methodology outlined in Chapter 4 to a set of 37 transcripts of the FDA Circulatory Systems Devices Advisory Panel Meetings. Implications of these findings are discussed in Chapter 7. The next chapter presents a quantitative model that attempts to replicate *in silico* the empirical findings outlined here, with a goal of deepening our theoretical understanding if decision-making on committees of technical experts.

Chapter 6

MODEL DEFINITION AND INITIAL RESULTS

> *"The meaning of a representation can be nothing but a representation. In fact it is nothing but the representation itself conceived as stripped of irrelevant clothing. But this clothing can never be completely stripped off: it is only changed for something more diaphanous. So there is an infinite regression here. Finally the interpretant is nothing but another representation to which the torch of truth is handed along; and as representation, it has its interpretant again. Lo, another infinite series."*
>
> *– Charles Sanders Peirce (1934-48), Collected Papers, 1.339, on modeling*

This chapter examines the role of expertise and one way that it might generate some of the empirical results observed in Chapter 4. To this end, we present a computational model whose purpose is to explore theoretical bases for the kinds of results found in Chapter 4. If the model can reproduce these empirical results and provide a potential explanation for the observed data, the underlying theory provides one potential explanation for the observed data. The results presented in this chapter are preliminary and subject to future investigation. The model presented here relies heavily on the literature in social psychology and, in particular, is a modification of the DISCUSS model, presented by Stasser (1992).

DISCUSS was originally developed to explain the discrepancy between shared and unshared information first observed by Stasser and Titus (1985). Although successful at replicating and explaining these findings (Stasser 1992), DISCUSS was not used to explain later findings regarding the nature of expertise (Stasser et

al. 1995). Indeed, the original version of the model treats all decision-makers as equivalent except for their initial information distributions.

In addition to incorporating expertise our version of DISCUSS must also be adapted to the case of the FDA panels. Thus, we must more closely examine FDA panel procedures. In particular, we would like the model to capture salient elements of the FDA panel process. The model presented in this chapter largely focuses on stages 5-8 of the panel process described in Chapter 3 (panel questions and later). Stages 1-4 (including sponsor and FDA presentations) are considered to be information revelation stages and fit into a pre-discussion phase as will be shown below.

The model takes as input the following variables:

Table 19: Model Input Variables

Name	Type	Range
Number of Speakers	Integer	4 – 15
Device Complexity (Number of Topics)	Integer	10 – 30
Device Quality	Real	-1 – 1
Device Ambiguity	Real	0 – 1
Specialty Membership	Logical Matrix	0 or 1; matrix has as many rows as speakers and as many columns as specialties. There may be

		as many as 8 specialties.
Mean Breadth of Expertise	Real	0 – 1
Dispersion in Breadth of Expertise	Real	0 – 1
Mean Depth of Expertise	Real	0 – 1
Dispersion in Depth of Expertise	Real	0 – 1
Speaker Hierarchy	Real	0 – 1
Process Openness	Real	0 – 1

The purpose of the model is to generate a simulated discourse between panel members – this discourse is then used to generate sample networks, whose properties can be compared to the data shown in Chapter 5.

Figure 70 shows a schematic of the model:

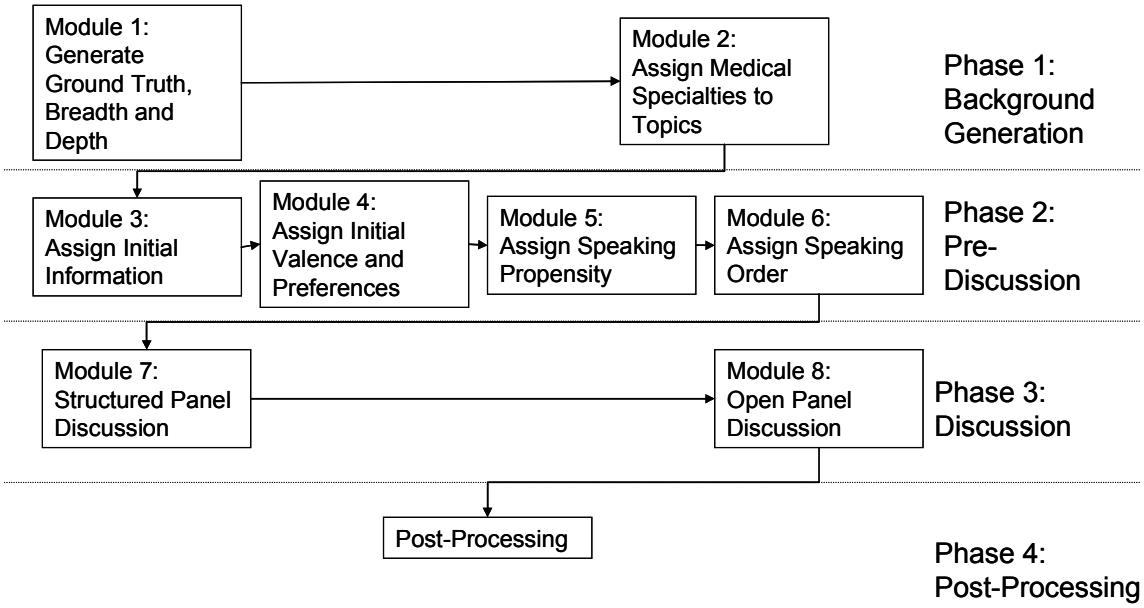


Figure 70: Schematic of the model outlined in this chapter.

As shown above, the model may be divided into four phases, which are then further subdivided into modules.

Phase 1: Background Generation

In this phase, conditions of "empirical reality" are determined based upon model input. Properties of the device and its associated domains of knowledge and expertise are generated from model inputs.

Module 1: Generate Ground Truth, Breadth and Depth

In this module, properties of the device are generated from the summary statistics used as input. In particular, we conceive of a device as having a finite set of features, or *topics*, which might describe it. The number of topics, n, required to describe a given device is equal to its *complexity* as defined in Table 19. Consistent with the DISCUSS model, these topics may be thought of as items of information that are necessary to fully describe whether a given device should or

should not be approved. Although there are formal problems with this assumption (Watanabe 1985), we consider it to be sufficient for modeling purposes. Future work might circumvent these concerns by introducing a structure relating topics to one another (Richards 2008). Each one of m *specialti*es has a different *perspective* on a topic, which is recorded as an entry in an a n x m matrix. Each perspective is assumed to embody information that is either pro- or anti- device approval. The intuition is that a device has a given *quality* ranging between -1 and 1, which describes, overall, whether it should or should not be approved. Furthermore, the data describing that device has an associated amount of *ambiguity* such that if the data is very ambiguous, different topics will give very different signals, whereas if the data is very unambiguous, different topics will give similar signals. For each feature, we would like to generate a number between -1 and 1 that describes whether and how strongly its associated topic supports or opposes device approval. We therefore construct a probability distribution that meets these requirements.

To do this, we use a beta distribution, which is defined by two parameters, α and β, and has the following form for its probability distribution function:

$$f(x;\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} \quad \textbf{(13)}$$

As can be seen from this expression, the range of a beta distribution is between 0 and 1. Furthermore, the mean of a beta distribution is given by the expression $\alpha/(\alpha+\beta)$. We would like to scale this range to be between -1 and 1. Therefore, quality, as defined above, is given by the following expression:

$$Q = \frac{\alpha}{2(\alpha+\beta)} + 1 \quad \textbf{(14)}$$

Further examination of the form of the beta distribution shows that for values of α and β less than 1, the beta distribution resembles an inverted-U, whereas for values of α and β near 1, the beta distribution resembles a uniform distribution. As α and β increase, the distribution has an increasingly large peak around the mean value. We take advantage of this property of the beta distribution to operationalize data ambiguity (A), which is defined on the interval 0 – 1. We first transform this number to A_t using the following formula:

$$A_t = e^{\frac{1}{A}-1} \quad \textbf{(15)}$$

This generates a number between one and infinity. We then use this effect to capture ambiguity as follows:

$$A_t = \alpha + \beta \quad \textbf{(16)}$$

Therefore as ambiguity goes to one, A_t goes to one, leading to a widely-dispersed distribution. Similarly, as A goes to 0, A_t goes to infinity, tending towards a distribution that is tightly concentrated around the mean. Solving these two equations simultaneously, we can determine values for α and β as follows:

$$\alpha = Q\, A_t \quad \textbf{(17)}$$

$$\beta = A_t - \alpha \quad \textbf{(18)}$$

Given these values of α and β, we then generate n random draws from a beta(α,β) distribution. Finally, these values are scaled onto the interval -1 – 1, defining a Ground Truth value for each topic/specialty pair, that fits the criteria described above. In particular, we define a Ground Truth matrix, **GT**, having m rows and n columns, such that for specialty i and topic j, $GT_{i,j} \sim$ beta(α,β).

Ground Truth is then re-scaled such that it is between -1 and 1, by multiplying each entry by two and subtracting one.

A similar method is used to assign values of breadth and depth to each speaker. In particular, each of these quantities is drawn from beta distributions with the parameters defined in Table 19 (Mean Depth/Breadth; and Dispersion in Depth/Breadth). With the exception of the re-scaling procedures used on the quality parameter and Ground Truth values, the transformations of the depth and breadth parameters are the same as those used for Ground Truth (e.g., $\alpha_{breadth}$=Mean_Breadth * Breadth_Dispersion$_t$; $\beta_{breadth}$= Breadth_Dispersion$_t$ - $\alpha_{breadth}$, where Breadth_Dispersion$_t$ is the result of the application of the transform in Equation 15 to the Breadth Dispersion shown in Table 19).

Module 2: Assign Medical Specialties to Topics

Each specialty is assigned a number of topics, representing information that members of that specialty are capable of knowing *a priori*. Each topic is assigned to each specialty with probability equal to 1/n. In addition, each topic is sequentially assigned to a given specialty in a deterministic fashion, such that the set of all topics is assigned sequentially among specialties. For example, if there are five topics and three specialties, then topic 1 is sequentially assigned to specialty 1, topic 2 to specialty 2, topic 3 to specialty 3, topic 4 to specialty 1, and topic 5 to specialty 2. As a result, each topic will be assigned to at least one specialty, and possibly more. Since each speaker is also assigned to a medical specialty, this forms the basis for an initial information distribution.

Phase 2: Pre-Discussion Phase

In this phase, preliminary information about each voting member is assigned prior to discussion.

Module 3: Assign Initial Information

A given speaker knows a given topic in his/her specialty in equal proportion to that speaker's depth. Recall that depth for each speaker is generated from a beta distribution with mean and dispersion parameters as shown in Table 19 and as transformed above. Furthermore, any speaker is capable of expertise. A speaker does not start with any knowledge about a topic outside this specialty.

Module 4: Assign Initial Valence

Each speaker is assigned a valence value for each topic about which s/he is knowledgeable. We may think of this value as the cognitive salience of that topic for that speaker, encoding its support or opposition to device approval. A beta distribution is generated for each author-topic-specialty triple. For author a seeing topic t from the perspective of specialty s, the mean of this distribution is the absolute value of Ground Truth – $|GT_{s,t}|$. The dispersion, d, is a function of author a's depth, such that $d = e^{\frac{1}{1-depth(a)}-1}$ as per equation 15. Thus, as depth increases, that speaker's initial valence is more likely to be close to Ground Truth. Each speaker's preference is the sign of the sum of their valences across all topics/specialty pairs in which they are knowledgeable.

Module 5: Assign Speaking Propensity

As in DISCUSS, speakers are assumed to generate an utterance with probability proportional to their location in a speaking hierarchy (cf. Stephan and Mishler 1952). In the DISCUSS model, if a speaker at the top of the hierarchy speaks with probability p, then the second speaker speaks with probability kp, the third with probability k^2p, etc, where k is the Speaker Hierarchy value defined in Table 19. Unlike the DISCUSS model, we assume the existence of two "lead reviewers". These are the first two speakers in the hierarchy, both of whom speak an equivalent amount, and three times more frequently than all other speakers. The lead reviewers each generate a number of utterances proportional to device

complexity and ambiguity throughout the meeting. In particular, the number of utterances for a lead reviewer is equal to 4 * Complexity * Ambiguity. Given that complexity ranges between 10 and 30, this allows the total number of utterances to go as high as 120, a number that is approximately equal to the total number of utterances empirically observed in the longest meetings. The remaining speakers follow the hierarchy defined in DISCUSS.

Module 6: Determine Speaking Order

In FDA panel meetings, speakers ask questions in a fixed order, followed by a period of open discussion. This order is determined by the direction in which the chair decides to go around the table in soliciting panel questions, and is therefore jointly dependent on seating location and the chair's procedural choice. Chwe (2003) notes that this is a form of ritual common knowledge. Lead reviewers always speak first, followed by the remaining panel members. This speaking order operates in the first part of the discussion phase, below.

Phase 3: Pre-Discussion Phase

The discussion phase consists of two modules: Structured Panel Discussion and Open Panel Discussion. The length of each of these depends on the Openness parameter shown in Table 19. Openness is defined as the maximum number of utterances in the Open Panel Discussion phase divided by the maximum number of utterances in the Structured Panel Discussion phase.

Module 7: Structured Panel Discussion

In this module, each speaker sequentially generates a finite set of utterances. The topic of each utterance is chosen with probability proportional to the absolute value of its valence, summed across all specialty perspectives with which that author is familiar for that topic. As in DISCUSS, we could allow for an advocacy parameter which might bias discussion of topics to those that support a given speaker's current preference. In particular, we can assume that speakers only

discuss those topics that support their current preferences. The benchmark model implemented below does not make this assumption.

Each utterance has the potential ability to influence other panel members to change their salience. Once a topic is discussed, each listener evaluates whether to adopt that topic's valence. With a finite probability, the listener adopts the salience of the speaker in that topic, t. The probability that a given listener will adopt the salience of a given speaker, a, is inspired by Latane's Social Impact Theory (Latane 1981), as follows:

$$P(adoption \mid t,a) \propto \frac{B_a * U_{t,a}}{(GroundTruth(t) - valence_{t,a})^2} \quad \textbf{(19)}$$

where $U_{t,a}$ is the number of utterances spoken by speaker a in topic t and B_a is a bias parameter. If the speaker and the listener share the same specialty, then B is equal to unity. Otherwise, B_a is equal to the breadth of the speaker. A speaker will adopt a topic's valence in proportion to the number of times that speaker has mentioned that topic (i.e., the speaker's perceived expertise in that topic) and in inverse proportion to the square of the distance of that topic's valence from Ground Truth (i.e., the speaker's actual expertise in that topic). The definition of perceived expertise presented here is limited to topic-specific air-time, as inspired by Bottger (1984). Future work might incorporate a notion of perceived expertise that changes members' contributions are judged valuable by others (such as panel members in other specialties).

Normalizing over all speakers yields the probability that a given listener adopts the valence of a given speaker. Preferences are then updated for all speakers. Otherwise, the listener retains his/her original valence value. This module terminates when all speakers have generated their assigned number of utterances.

Module 8: Open Panel Discussion

This module is equivalent in nature to the Structured Panel Discussion module with the exception that speakers are chosen at random in proportion to their speaking propensity as defined in Module 6. This module terminates if the panel has reached a near- (i.e., minority of size 1) or full-consensus, or if each speaker has used up their full set of utterances.

Phase 4: Post-processing

Once the discussion phase has terminated, networks are generated from the simulated discourse. Since, in this model, documents are assigned deterministically to topics, there is no need to create a distribution over networks for each meeting. Instead, two authors are linked if they both generate at least one utterance in the same topic.

Model Benchmarking

Although the model presented in this chapter is a sparse representation of the actual dynamics within FDA panels, we may explore whether its outputs reflect the data shown in Chapter 4. 1000 samples were drawn from the model, while allowing mean depth, mean breadth, dispersion in depth, dispersion in breadth and openness to vary. These parameters were all drawn from uniform distributions on the unit interval. Hierarchy was set to 0.99, consistent with panel procedures that attempt to allocate approximately equal amounts of time to all voting members. So as to enable a direct comparison, the distribution of specialists and the number of panel members were chosen according to the 37 meetings tested previously (i.e., one sample from the model would have the same number of panel members and the same distribution of specialties as a randomly chosen Circulatory Systems Devices Panel meeting).

As in Chapter 5 there is a tight correlation between simulated medical specialty cohesion and simulated vote cohesion (Spearman Rho = 0.57; $p=8.97 \times 10^{-16}$) for

the subset of meetings in which there was a voting minority consisting of at least two members. A scatter plot is shown in Figure 71.

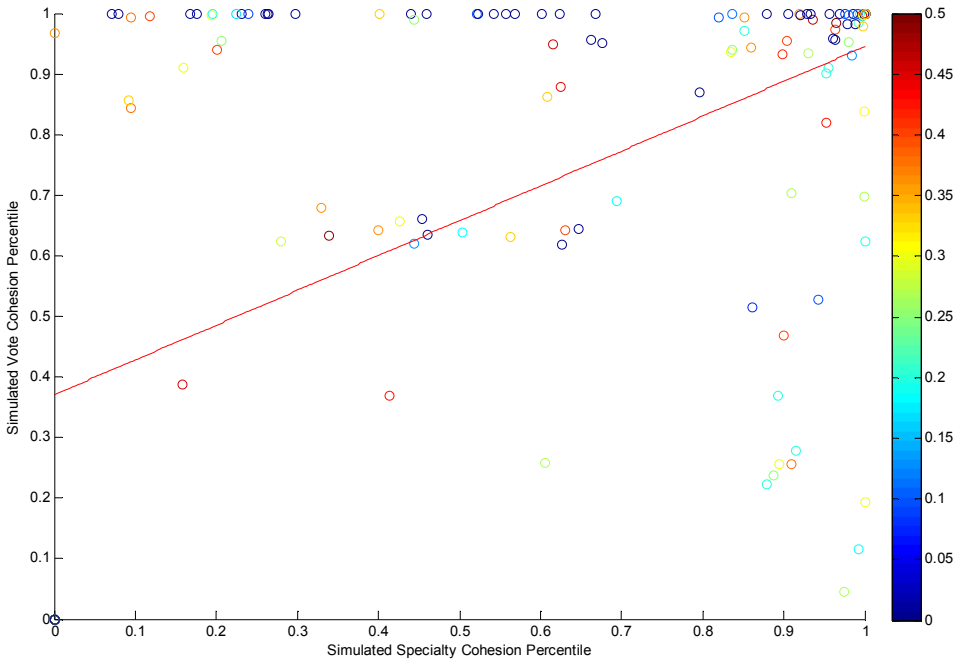


Figure 71: Plot of Simulated Specialty Cohesion vs. Simulated Vote Cohesion. Spearman Rho = 0.57; $p=8.97 \times 10^{-16}$. Proportional minority size (including abstentions) is represented in color.

Modeling Result 1: Simulated vote cohesion percentile and simulated specialty cohesion percentile are significantly positively associated.

As in Chapter 4, we find that members of the simulated voting minority tend to speak later than do members of the simulated voting majority ($p<0.0001$, see Figure 72).

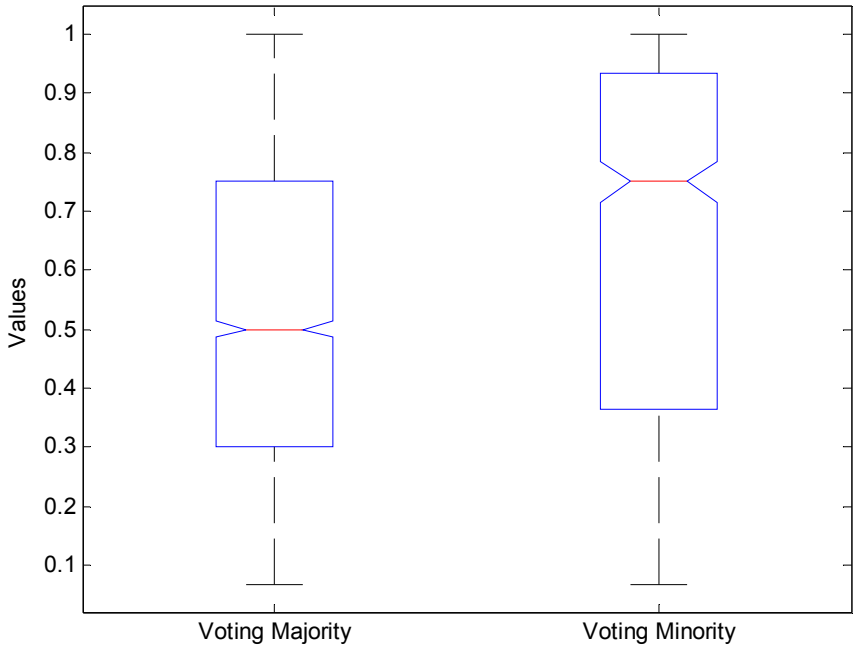


Figure 72: Members of the simulated voting minority tend to speak later than do members of the simulated voting majority ($p<0.0001$).

Another way in which we might verify the model fit is to examine the extent to which the specialty cohesion, specialty cohesion percentile, vote cohesion, and vote cohesion percentile distributions fit the observed data. Kolmogorov-Smirnov tests show that the model generates distributions that are not significantly different from those observed in the empirical data (see Table 20).

Table 20: Kolmogorov-Smirnov tests show no significant differences between the empirical and simulated distributions for vote cohesion and for specialty cohesion.

Distribution Tested	Probability that data is consistent with hypothesis of no significant difference (p-value)
Specialty Cohesion	0.68
Specialty Cohesion Percentile	0.12
Vote Cohesion	0.18
Vote Cohesion Percentile	0.11

One further test would be to examine the proportion of simulated meetings that reached consensus, those that have a voting minority consisting of just one member, and those that have a voting minority with more than one member. Recall that, in the empirical case, there were 11 meetings with at least two members in the voting minority, and there were 6 meetings with only one member in the voting minority, out of a total of 37 meetings. Table 21 shows that the empirical and simulated distributions are significantly different ($p=4.54 \times 10^{-5}$).

Table 21: A chi-square test shows that the simulated data distribution of panel meeting outcomes does not match the empirical distribution (χ^2 =20.00; dof=2; p=4.54 x 10^{-5})

	Panel consensus	One voting minority member	Larger voting minorities	TOTAL
Simulated Data	811	41	148	1000
Empirical Data	20	6	11	37
TOTAL	642	143	252	1037

In general, we find that there are larger voting minorities in the empirical data than in the simulated data. One possible explanation for this is that simulated panel members are not reviewing devices that are as uncertain. We correct for this by constraining device quality to vary between -0.33 and 0.33, ensuring that those devices reviewed in the simulation do not have any "easy answers". This is consistent with the role of the panels in reviewing only difficult devices. An additional 1000 samples from the model were drawn, yielding results that are consistent with the empirical data. Table 22 shows that under these conditions, the two distributions examined above are not significantly different.

Table 22: A chi-square test shows that the simulated data distribution of

panel meeting outcomes does not match the empirical distribution (χ^2 =3.02; dof=2; p=0.22)

	Panel consensus	One voting minority member	Larger voting minorities	TOTAL
Simulated Data	639	85	276	1000
Empirical Data	20	6	11	37
TOTAL	659	91	287	1037

Kolmogorov-Smirnov tests continue to show no significant difference in vote and specialty cohesion percentile distributions, as shown in Table 23.

Table 23: Kolmogorov-Smirnov tests show no significant differences between the empirical and simulated distributions for specialty and vote cohesion or for their percentiles.

Distribution Tested	Probability that data is consistent with hypothesis of no significant difference (p-value)
Specialty Cohesion	0.36

Specialty Cohesion Percentile	0.095
Vote Cohesion	0.91
Vote Cohesion Percentile	0.12

We also continue to observe that members of the voting minority speak later than do members of the voting majority ($p<0.0001$), and that there is a strong correlation between simulated voting cohesion percentile and simulated specialty cohesion percentile (Spearman Rho = 0.61; $p=1.70 \times 10^{-33}$).

These parameters seem to fit the data reasonably well, with the exception of a tendency for the simulated data to have larger values of specialty cohesion percentile and vote cohesion percentile than does the empirical data. Therefore we set this as the benchmark from which we can test the effects of deviation of other parameters. Future work will focus on determining a mechanism by which the simulated percentile distributions may better fit the empirical data.

Model Result 2: Proportional minority size, and specialty and vote cohesion, are functions of device quality.

Deviations from the Benchmarked Model

Random Speaking Order

We find that, in the absence of a pre-set speaking order, there is no significant difference between the locations in the speaking order of members of the voting majority and the voting minority ($p=0.69$, see Figure 73).

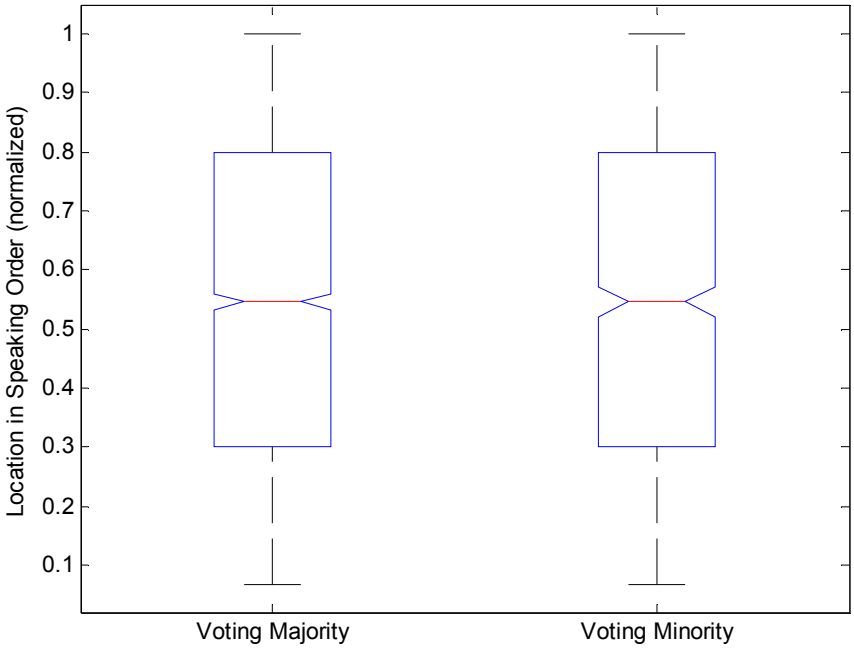


Figure 73: In the absence of a pre-set speaking order, members of the simulated voting minority do not tend to speak later than do members of the simulated voting majority (p=0.69).

Interestingly, when speaking order is randomized, Kolmogorov-Smirnov tests yield significant differences in vote and specialty cohesion percentiles when compared to empirical data (see Table 24).

Table 24: Kolmogorov-Smirnov tests show significant differences between the empirical and simulated distributions for specialty and vote cohesion percentiles.

Distribution Tested	Probability that data is consistent with hypothesis of no significant difference (p-value)
Specialty Cohesion	0.37
Specialty Cohesion Percentile	0.032
Vote Cohesion	0.88
Vote Cohesion Percentile	0.036

In addition there is a stronger correlation between simulated voting cohesion percentile and simulated specialty cohesion percentile than that observed under pre-set speaking order conditions (Spearman Rho = 0.10; $p=1.30 \times 10^{-41}$).

Modeling Result 3: Members of the simulated voting minority tend to speak later than do members of the simulated voting majority when only when speaking order is pre-set. Furthermore, simulated vote and specialty cohesion percentile distributions fit the data better when speaking order is pre-set. Correlation between specialty cohesion percentile and vote cohesion percentile is slightly stronger than under conditions of pre-set speaking order.

Advocacy

Assuming full advocacy – i.e., that panel members only discuss topics whose valences are consistent with their preferences, changes the model's results such that, under these conditions, the correlation between vote cohesion percentile and specialty cohesion percentile strengthens (Spearman Rho = 0.84; $p<0.0001$). In addition, the Kolmogorov-Smirnov tests show significant differences between empirical and simulated vote and specialty cohesion, and vote and specialty cohesion percentiles (see Table 25).

Table 25: Kolmogorov-Smirnov tests shows significant differences between the empirical and simulated distributions for vote cohesion and specialty and vote cohesion percentiles.

Distribution Tested	Probability that data is consistent with hypothesis of no significant difference (p-value)
Specialty Cohesion	0.87
Specialty Cohesion Percentile	1.01×10^{-5}
Vote Cohesion	2.36×10^{-5}
Vote Cohesion Percentile	0.041

Modeling Result 4: The model better explains the empirical data distributions when no advocacy among panel members is assumed,

although correlation between specialty cohesion percentile and vote cohesion percentile is stronger than under conditions of non-advocacy.

Expertise and Meeting Parameters

A 5-way Analysis of Variance, shown in Table 26, demonstrates that complexity, mean breadth, mean depth, depth dispersion, openness, and meeting profile (number and diversity of panel members) are all significant variables affecting specialty cohesion percentile.

Table 26: 6-way Analysis of Variance showing the impact of Complexity, Mean Breadth, Mean Depth, Depth Dispersion, Openness and Ambiguity on Specialty Cohesion Percentile

Variable	Sum of Squares	Degrees of Freedom	Mean Squares	F	p-value
Complexity	1.23	1	1.23	10.92	0.001
Mean Breadth	2.87	1	2.87	25.53	<0.0001
Mean Depth	5.39	1	5.39	47.99	<0.0001
Depth Dispersion	2.14	1	2.14	19.03	<0.0001
Openness	1.12	1	1.12	10.00	0.0016
Ambiguity	14.26	1	14.26	126.87	<0.0001

Error	116.60	994	0.11
Total	136.94	999	

Modeling Result 5: Simulated complexity, mean breadth, mean depth, depth dispersion, openness and ambiguity all have a significant effect on simulated specialty cohesion percentile.

We find that specialty cohesion percentile decreases with mean breadth and openness, and increases with mean depth, depth dispersion, complexity and ambiguity.

A 2-way ANOVA, shown in Table 27, shows that complexity and ambiguity are both significant predictors of voting cohesion percentile.

Table 27: 2-way Analysis of Variance showing the impact of Complexity and Ambiguity on Vote Cohesion Percentile

Variable	Sum of Squares	Degrees of Freedom	Mean Squares	F	p-value
Complexity	1.41	1	1.41	14.72	0.0002
Ambiguity	4.89	1	4.89	51.05	<0.0001
Error	30.18	315	0.096		
Total	37.01	317			

Modeling Result 6: Simulated complexity and ambiguity have a significant effect on simulated voting cohesion percentile.

Vote cohesion percentile increases with complexity and ambiguity.

Given that both specialty cohesion percentile and vote cohesion percentile depend on ambiguity and complexity, we may use this information to further fine-tune the model. For example, an analysis of covariance (see Figure 74) shows that the correlation between specialty cohesion percentile and vote cohesion percentile is significantly weaker when ambiguity is greater than 0.5 ($p=1.26x10^{-5}$).

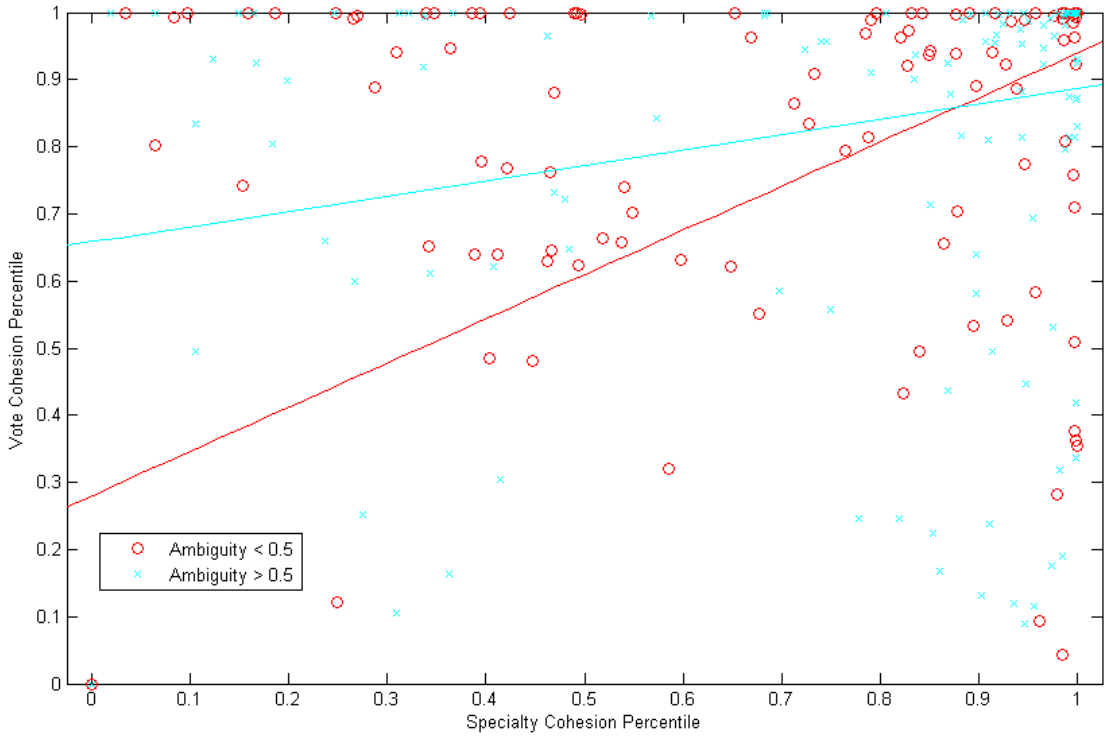


Figure 74: Analysis of Covariance Plot showing the effect of ambiguity on correlation between Specialty

Cohesion Percentile and Vote Cohesion Percentile

A similar analysis shows that the correlation between specialty cohesion percentile and vote cohesion percentile is significantly weaker when complexity is greater than 20 topics (p=0.0078, see Figure 75).

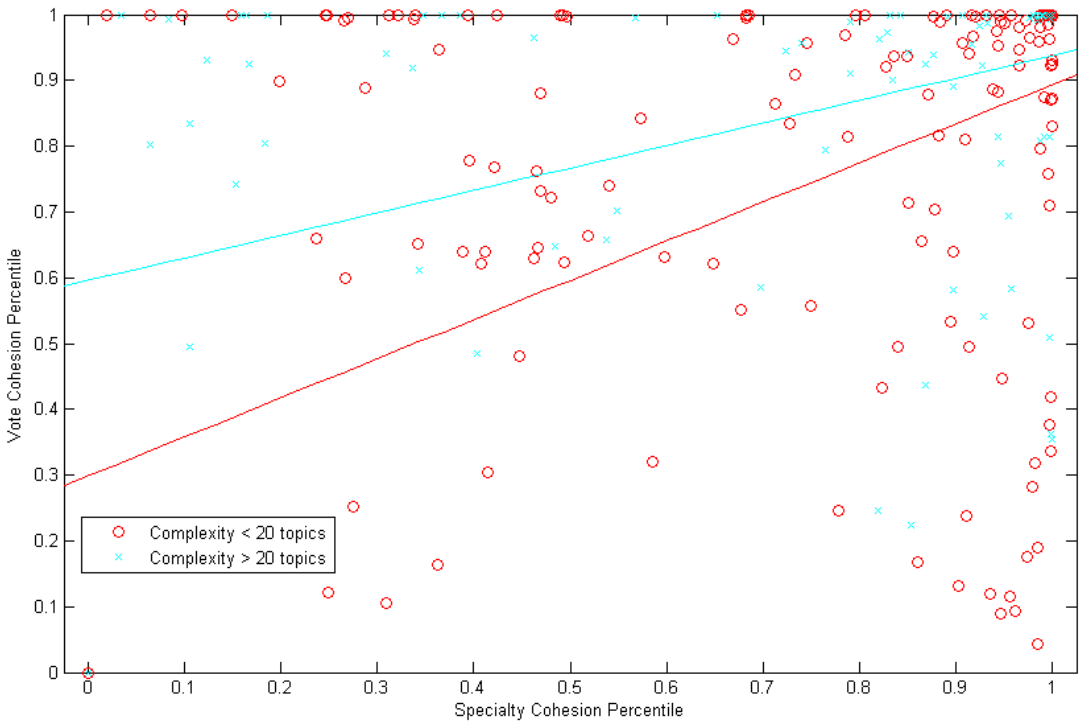


Figure 75: Analysis of Covariance Plot showing the effect of complexity on correlation between Specialty Cohesion Percentile and Vote Cohesion Percentile

Modeling Result 7: Association between specialty cohesion percentile and vote cohesion percentile is stronger when ambiguity is less than 0.5 and when complexity is less than 20 topics.

Panel's Ability to Reach Consensus

We have noted above that the panel's ability to reach consensus depends on device quality. Other factors include diversity, complexity, mean breadth, and ambiguity, all of which are significant predictors of proportional minority size (see Table 29).

Table 28: 5-way Analysis of Variance showing the impact of Diversity, Complexity, Mean Breadth, Mean Breadth, Quality and Ambiguity on proportional minority size.

Variable	Sum of Squares	Degrees of Freedom	Mean Squares	F	p-value
Diversity	0.095	1	0.095	5.66	0.018
Complexity	0.20	1	0.20	12.05	0.0005
Mean Breadth	0.28	1	0.28	16.79	<0.0001
Quality	0.79	1	0.79	47.03	<0.0001
Ambiguity	0.71	1	0.71	42.12	<0.0001
Error	16.69	994	0.017		
Total	18.89	999			

In particular, proportional minority size increases with diversity and ambiguity, and decreases with complexity, mean breadth, and quality.

Modeling Result 8: Simulated proportional minority size is significantly associated with diversity, complexity, mean breadth, quality and ambiguity.

Panel's Ability to Correctly Decide

The simulated panel correctly decided the correct outcome in 778 of the 1000 meeting samples drawn from the model. As expected, mean depth, quality and ambiguity are all significant predictors of whether the voting outcome is correct (see Table 29).

Table 29: 5-way Analysis of Variance showing the impact of Complexity, Mean Depth, Openness, Ambiguity and Quality on correct vote outcome. Although correct vote outcome is a dichotomous variable, the analysis is still qualitatively instructive (Lunney 1970).

Variable	Sum of Squares	Degrees of Freedom	Mean Squares	F	p-value
Meeting Profile	8.38	36	0.23	1.55	0.021
Complexity	1.71	1	1.71	11.38	0.0008
Mean Depth	2.60	1	2.60	17.35	<0.0001

Depth Dispersion	1.02	1	1.02	6.8	0.0093
Openness	0.68	1	0.68	4.55	0.033
Ambiguity	4.62	1	4.62	30.82	<0.0001
Quality	8.19	1	8.19	54.67	<0.0001
Error	143.42	957	0.15		
Total	172.72	999			

Modeling Result 9: Simulated meeting profile, complexity, mean depth, depth dispersion, openness, ambiguity and quality are all significantly associated with the panel's capacity to reach a correct decision.

In particular, as complexity, depth, depth dispersion, and quality increase (or move away from zero, in the case of quality), the panel is more likely to reach a correct decision, whereas as ambiguity and openness increase, the panel is less likely to make a correct decision. The significant effect of meeting profile may be understood as a control variable. In some configurations, the panel always generated the right answer (e.g., meetings 5 and 29), whereas in others, the panel was often incorrect (e.g., meeting 36, in which the panel was correct only 60% of the time). The interesting question of why these profiles yielded these outcomes is left to future work.

Preliminary Modeling Conclusions and Future Work

The model presented in this chapter is still in its infancy, and fails to explain many facets of the operations of FDA panels. Nevertheless, we may derive many insights from this analysis. In particular, the fit of the model to some aspects of the data, even using randomized parameters, suggests that some major elements on FDA panels are captured. These, and their implications, are discussed in Chapter 7.

C h a p t e r 7

CONCLUSIONS

"...natural languages are perfect in so far as they are many, for the truth is many-sided and falsity consists in reducing this plurality into a single definite unity."

– Umberto Eco, The Search for the Perfect Language, (1997), trans. Italian. James Fentress, on the benefits of diversity

Committees of experts are essential to engineering systems because of their ability to aggregate information from multiple domains of expertise – a necessary function when system complexity is large. This has the potential to enable better decision-making about a complex problem. Assuming a perfect flow of communication, we can expect diverse committees to pool their knowledge to make far better decisions than individuals possible with the information available (Hong & Page 2004). Even with less than perfect information flow, improved outcomes over individuals are likely. The literature indicates that communication flows on these committees are crucial to optimal decision-making. In particular, committee members with the appropriate expertise must be given the opportunity to express their views (Bottger 1984). Furthermore, their advice must be appropriately received and interpreted by the majority of committee members. The literature suggests that the correct procedural interventions could create conditions under which communication for information sharing could be optimized. In particular, if at all possible, members with appropriate expertise must be identified as such early in the decision-making process (e.g, through the assignment of lead reviewer status). In the case that there is a disagreement

regarding whose expertise is valid, the literature suggests that a decision might be made on other bases, with committee members potentially defaulting to other decision-making schemes, such as those controlled by idiosyncratic beliefs and values. Furthermore, different perspectives might be incommensurable, leading to a persistent disagreement and a split-vote on the panel. The literature suggests that, in the presence of clear, unambiguous data, role- and preference ambiguity inherent in the decision-making process is reduced since panel members would likely be able to agree on the interpretation of a significantly well-defined data artifact.

Better understanding the role of information flow on committees of technical experts requires a methodology that can used to study real-world committee decision-making. The method developed in this thesis is based upon the Author-Topic model (Rosen-Zvi et al. 2004), and is able to extract meaningful directed social networks from transcripts of the FDA Circulatory Systems Devices Panel Meetings. These networks represent the flow of communication, and potentially information, among panel members.

Empirical Results and Their Implications

Analysis of the networks generated from the methodology described in Chapter 4 has yielded 27 empirical results, with the following implications:

Air-Time on FDA Panels

1. Gender, Medical Specialty, h-Index, and Age are all significant variables associated with a panel member's air-time. Women tend to have more air-time than men do, although this effect is not visible in meetings with voting differences.

Although there is an effect for gender on air-time, it goes in a direction opposite than that predicted by (Berger et al. 1972), with female speakers using more air-

time than do male speakers, and seems to disappear entirely for meetings in which there is a voting-difference present. Air-time increases with h-index (Spearman's Rho=0.26). After correcting for h-index variation, air-time decreases with age, although the effect is very weak (Spearman's Rho=-0.06). Finally, when controlling for the variables outlined above, we find no effect of race on air-time. These results suggest that FDA panel procedures are largely free from the status effects predicted in the "small-groups" sociology literature. One possible explanation of this may be found in Festinger's (1954) theory. In the presence of a clear task requiring expertise, generalized status effects might become less important with panel members instead focusing perceived expertise on measures that are directly relevant to the task at hand (e.g., h-index and other metrics of academic or clinical experience).

2. There is no observably significant effect between vote and air-time.

This implies that the FDA process seems to allow those in the voting minority an equivalent amount of time to speak as those in the voting majority. Viewed within the context of Bottger's (1984) distinction between perceived influence (based on air time) and actual influence (based on expertise) we see that the two are quite independent. Bottger predicts that performance should increase as expertise and air-time covary; however, given that all FDA panel members are acknowledged experts in their respective fields, this lack of covariance is not surprising and might be attributed to a procedure whose goal is to ensure that many different, but valid, viewpoints are heard in a public forum. Indeed, these findings suggest that in a structured task, such as on FDA panels, perceived expertise may not be equivalent to air time, For example, *a priori perceived expertise* might be explained by a procedural variable (e.g., speaking order). Under such circumstances, where each panel members has an opportunity to talk, *a posteriori* perception of expertise could match actual expertise.

3. There is no observably significant effect of medical specialty, h-Index, age or race on voting behavior. Women are more likely to be in the voting majority than men are.

The prevalence of women in the voting majority, coupled with the additional air-time used by women when consensus meetings are included may be due to the role of women as information integrators on committees (Johnson & Eagly 1990), and suggests future work in determining the balance on a committee between broad integrators versus deep specialists.

Empirical findings 1 - 3 seem to suggest that FDA panel procedures are successful in avoiding both perceived and actual bias associated with commonly expected attribute-based status characteristics. The absence of a difference in air-time between majority and minority members further contributes to the perception of the panel meeting as a fair and balanced process, in which minority and majority members may equally express their views. It is particularly interesting that there is no significant impact of h-index on vote ($p=0.66$, using a Kruskal-Wallis non-parametric one-way ANOVA), which could be explained in that the FDA panel process might weight academic publication record against other sources of expertise, such as clinical experience.

Medical Specialty as a Mediating Variable

4. There is no observably significant effect between medical specialty and vote.

Medical specialty, and technical training in general, is not a status characteristic as defined in the sociology literature. Specialties are different areas from which a panel may draw upon a diversity of expertise. Associated with these specialties are different standards for evaluating expertise and different speaking habits (cf. empirical finding 1). Social scientists might term these as different "professional

cultures" or "institutions" (e.g., Douglas 1986, Trice 1993). Empirical finding 4 suggests that medical specialty is a cross-cutting categorization that influences behavior in a more subtle way.

5. Panel members of the same medical specialty are significantly more likely to be linked than would be expected by chance.

(Brown 1986; Douglas 1986) note that members of a common professional institution are likely to share common language and jargon. This may explain empirical finding 5, which finds that panel members with the same medical specialty tend to be linguistically linked. This indicates that the majority of communication on most FDA panels likely occurs between members of the same medical specialty. Nevertheless, the presence of some probability mass that is less than or equal to 0.5 may indicate meetings where some panel members made stronger attempts to communicate across specialty boundaries.

6. Panel members who vote the same way are significantly more likely to be linked than would be expected by chance.

This finding suggests that panel members who vote the same way share the same language. One possible explanation of this phenomenon is that these panel members are focusing their attention on a common area, such as a component of the device or an aspect of the sponsor's data. One strategic interpretation is that different panel members may have similar preferences *a priori*, and would therefore focus on a device's common shortcomings or merits to signal their preferences. It is more likely that within each voting group, a relatively small number of device features might attract the attention of a number of panel members, causing them to vote a certain way for that reason. Common language could suggest a common direction of attention and therefore, common

preferences. This might arise as panel members successfully learn from one another.

7. Vote cohesion percentile and specialty cohesion percentile are significantly positively associated for the subset of 11 meetings with at least two members in the voting minority.

In cases of mild ambiguity, where a small number of potential interpretations of the data are possible, (Douglas 1986) notes that institutional membership acts to direct one's attention to a given framing of a situation or problem. This framing mechanism could potentially serve as an antecedent to preference formation. If such is the case, then a correlation between vote cohesion percentile and specialty cohesion percentile would be expected. We may use this insight to explain empirical finding 7 by assuming that a medical specialty directs a given voter's attention to a certain interpretation of the data, thereby creating conditions under which members of a given medical specialty will pay attention to the same things. Within the medical community, Kaptchuk (2003) calls this phenomenon "interpretive bias". This common perception of the data leads to a propensity to vote in a manner consistent with that perception. This is further supported by the fact that when specialty cohesion is low, voting cohesion also tends to be low. In these situations, it is likely that the data is difficult to interpret, e.g., due to mixed signals from a device that has a high risk but high potential reward, or sparse or ambiguous data. Under such conditions, many possible interpretations of the data might be possible within each specialty, suggesting that voters could rely on idiosyncratic beliefs. Medical specialties would have a weaker effect on an individual's perception since the data might not match any situation previously encountered. Specialty cohesion would be lower because panel members from the same specialty would have different perceptions of the data. Under these circumstances, individual expertise becomes particularly valuable, although it is

unclear whose expertise is most appropriate. Panel members who vote the same way would likely do so for different reasons, thus leading to low vote cohesion. This finding cannot account for those meetings in which the panel reached consensus or only had one member in the voting minority. In these cases, voting cohesion has no meaning, whereas specialty cohesion runs the gamut of values. Low specialty cohesion during a consensus meeting might indicate learning across medical specialty boundaries, ultimately leading to a common interpretation.

Agenda-Setting and the Effects of Speaking Order

8. Members of the voting minority tend to speak later than do members of the voting majority.

One interpretation of this finding might be the presence of framing and agenda-setting effects, such as identified in (Cobb & Elder 1983). This would seem to suggest that those panel members who speak first are more likely to influence other panel members, because later speakers must respond to the problem as it has already been framed by the first speakers. Later speakers are less likely to have influence over defining the issues discussed in the panel meeting and their opinions are therefore more likely to be in the minority. This suggests the possibility that vote might be influenced by what have been called "ritual" elements in the literature (e.g. Chwe 2003; Douglas and Wildavsky 1982). In particular, speaking order typically begins with the lead reviewers and then proceeds sequentially around the table in a direction chosen by the committee chair. Choice of seating location is jointly determined before the meeting by the committee chair and FDA executive secretary (FDA 1994).

9. Members of the voting minority tend to have a lower graph outdegree than do members of the majority.

10. Outdegree is negatively and significantly associated with location in the speaking order, and indegree is positively and significantly associated with location in the speaking order.

11. Location in the speaking order seems to account for the variance in voting behavior that is associated with outdegree.

Taken together, empirical findings 9-11 suggest that panel members who speak later are less likely to be repeated even in the presence of multiple rounds of discussion. If empirical finding 9 suggests the presence of an order-based hierarchy, then empirical finding 10 finds that it extends beyond the simple speaking order of the first round of panel questioning and throughout the meeting. All of the information above points to the role of speaking order as an important procedural variable, potentially embodying a form of perceived expertise. The FDA Policy and Guidance Handbook (1994) emphasizes the role of the FDA Executive Secretary and the committee chair in choosing the seating order of different panel members, suggesting that this is one possible lever by which control over the decision-making process could potentially be exercised. Ideally, seating order would correlate with actual expertise. This is often the case when expert lead reviewers are chosen to speak first. In other cases, it may not be clear *a priori*, which expertise is most relevant. In these situations, it is important to be aware of the potential procedural consequences of seating order.

12. Members of the voting minority spoke significantly later in meetings in which the panel did not approve the devices than did members of the voting majority. This trend was not present in meetings in which the panel did approve the device.

One possible explanation of this effect might be that there is not enough statistical power to differentiate between voting minority and voting majority

groups due to the smaller number of meetings in which there was a voting minority and the device was approved. Another possible explanation might be that negative comments about a device are weighted more strongly than are positive comments as per Kahneman and Tversky's (1979) Prospect Theoretic bias in which "losses loom larger than gains". If negative opinions are expressed relatively early in the panel discussion, this might predispose the panel to vote against device approval. If such is the case, then one way to counteract such a bias might be choose speaking order to insure that members who are likely to contributed negative comments speak later[15]. This would best enable both positive and negative comments to be expressed, ensuring a balanced process. More data is necessary to test this hypothesis.

13. Members of the voting minority had a significantly smaller outdegree in meetings in which the device was approved. This trend was not present in meetings in which the panel did not approve the device.

This suggests that, in meetings in which the device was approved, members of the voting minority seemed to exercise less influence over topic selection than did members of the voting majority. On the other hand, in meetings in which the device was not approved, members of the voting majority and the voting minority tend to have the same amount of influence.

Together, empirical results 12 and 13 suggest that when there is a minority and devices are approved, the voting minority (i.e., those who voted against the device) has less influence over topic selection and so may not need to be located later in the speaking order. On the other hand, when there is a minority and devices are not approved, the voting minority (i.e., those who voted in favor of the device) have more influence over topic selection despite their location later in

[15] The author would like to thank Dr. Susan Winter for suggesting this interpretation.

the speaking order. One might simply interpret this data in terms of a disparity between perceived and actual expertise. We would expect that minority panel members would have less actual influence (and hence, a lower outdegree) in a panel meeting in which the majority and the minority are equally distributed throughout the speaking order (i.e., the procedure does not embody perceived expertise). On the other hand, when procedural effects seem to locate the voting minority at the end of the speaking order (as is the case in non-approval meetings), we notice that the actual influence of the majority and the minority are statistically indistinguishable. In such cases, it might be that perceived expertise and actual influence do not covary.

Another interpretation would seem to suggest an opposite effect to that discussed after empirical finding 12. Those who are in favor of device approval seem to have an outdegree that is at least as high as those who oppose device approval, even when they are located at the end of the speaking order. On the other hand, those who are opposed to device approval seem to have a lower outdegree, even when they are not significantly later in the speaking order. One way of interpreting this result is that it might suggest a predisposition towards approval on FDA panels, particularly since panel members might choose to impose conditions of approval rather than rejecting the device wholesale. Examining the direction of causation underlying these effects is an area for future work.

14. Members of the voting minority are more likely to be graph sinks than are members of the voting majority.

Even in situations in which voting minority members aren't linked to one another, they are still likely to be graph sinks. This suggests that independent meaning may be derived from graph sink status. For example, voting minority members may not agree with the voting majority for different reasons which other voting members do not repeat.

15. Members of the voting minority are more likely to be the last speaker to ask questions to the sponsor and FDA, than are members of the voting majority, especially for meetings in which there is a singleton voting minority.

16. Using F-score as an evaluation criterion, the graph sink heuristic provides a superior classification of voting minority members when compared to the last speaker heuristic.

Empirical findings 12 and 13 follow as consequences of the effects of speaking order. On the other hand, empirical finding 14 shows that introducing topic-related information can aid in classifying minority members, and may point to a dynamic on the panels wherein members of the voting minority may be unable to convince other panel members to adopt their perspectives. The examination of graph sinks is a better method for determining voting minority membership than is examining the last speaker. In the case of the 17 meetings with a minority, precision is higher using the graph sink heuristic, suggesting that graph sinks capture a higher proportion of minority members than do last speakers. This makes sense given that there can be multiple sinks per meeting, but only one last speaker. On the other hand, recall is higher using the last speaker heuristic, suggesting that a randomly chosen member of the minority is more likely to be correctly classified using the last speaker method. This can be explained by the fact that a meeting with only one minority member might have multiple sinks, introducing a potential source of noise, perhaps due to valence effects (i.e., the tendency not to oppose the majority for reasons of maintaining one's reputation – cf. Ashworth & Bueno de Mesquita, 2005). These additional sinks might not be influential, even though they vote with the majority. We find that F-score is higher for the graph sink heuristic, suggesting that the noise in the recall metric is more than offset by the advantages gained in the precision metric. This

conclusion is stronger in the subset of 11 meetings with a voting minority of 2 or more, where recall is also higher using the graph sink metric. This suggests that, for larger minorities, the graph sink metric becomes increasingly accurate in classifying voters.

Framing and Ambiguity

17. Although lead reviewers have a significantly higher air-time and outdegree, and a significantly lower indegree than other panel members, their overall voting behavior is not significantly different.

This result is surprising in light of the effects of speaking order identified above and seems to contradict a major tenet of the agenda-setting literature – namely that those with the capacity to frame an issue can set the agenda and, therefore, strongly influence decision outcomes. One would expect lead reviewers to be more frequently in the majority; nevertheless, the lead reviewers' probability of being in the majority is statistically indistinguishable from that of the rest of the panel (the values of χ^2 are near zero, and the p-values are close to 1). One possible explanation of this result is that the FDA might choose lead reviewers that are representative of different perspectives on a given device, with the intention of fostering open communication. Thus, when lead reviewers disagree, it might be reflective of a wider split within the medical community. On the other hand, it might be that the panel's vote distribution follows that of the lead reviewers present. We cannot determine the direction of causation from this analysis – nevertheless, it is clear that the votes of the lead reviewers are indeed correlated with the votes of the rest of the panel in some manner.

18. The proportional size of the voting minority is larger when lead reviewers do not vote with the majority within a given meeting, and when there is disagreement among lead reviewers.

As stated above, the direction of causation underlying this result is unclear. One possible explanation, following the agenda-setting literature, is that there are multiple competing interpretations in situations in which lead reviewers disagree – each frame is a different description of the problem and leads to different conclusions regarding the appropriateness of approving the device. Another explanation is that the device or the data describing it is inherently ambiguous and that there is deep uncertainty regarding its suitability for approval. These two explanations are not necessarily contradictory, since the existence of multiple credible frames is much more likely as ambiguity increases (cf. March 1994). The increase in the proportional size of the majority may therefore be a reflection of this ambiguity as voting members preferentially adopt frames. Often, these are related to an individual's medical specialty, which serves to direct that individual's attention to a particular set of salient device characteristics.

19. Meetings are longer when more lead reviewers are in the voting minority.

One possible interpretation of this result is that as meeting length increases, the procedural components of the meeting become relatively less important – i.e., there is more open discussion. This might happen because of disagreement regarding how to interpret a given set of clinical trial data, or because of some other source of ambiguity (March 1994). The presence of multiple competing interpretations, possibly advanced by the presence of lead reviewers who disagree with other prominent panel members (or other lead reviewers) supports the notion that that ambiguity is a driver of dissensus on FDA panels. As the panel works to reconcile these different perspectives, meetings may take more time.

20. Meeting length is significantly positively associated with the maximum normalized outdegree among voting minority members, but not with maximum location in the speaking order.

As meetings get longer, more time is likely to be devoted to open, non-structured discussion. The hierarchy established by speaking order is therefore less likely to have as strong an impact on voting outcome. Furthermore, in some longer meetings, lead reviewers may be more likely to be in the voting minority. Thus, as meetings get longer, we see voting minority members begin to appear higher on the graph and to have a larger outdegree.

21. Maximum normalized outdegree is significantly associated with proportional voting minority size.

22. Meeting length is significantly positively associated with proportional voting minority size.

Empirical results 17-22 show a set of four variables that are mutually correlated: meeting length, proportional voting minority size, maximum voting minority outdegree, and minority voting dissensus. Such clusters of correlated variables are indicative of a "natural mode" (Richards 2008b), and might suggest the presence of an underlying driver. One explanation is the presence of device ambiguity – as the data regarding a particular device becomes harder to read, it leads to the possibility for multiple possible interpretations of the data. These interpretations are often mutually inconsistent, and may require additional time to resolve. Alternatively, they might not be resolved at all, leading to a split vote on the panel. When different interpretations from different experts are equally probable, it is possible that lead reviewers will be chosen who reflect this dichotomy. Nevertheless, it is precisely under conditions of ambiguity that different interpretations, based upon different values, may dominate (March 1994).

Impact of the Committee Chair

23. Inclusion of the Committee Chair in directed graphs leads to a bimodal distribution of the extent to which the chair changes the structure of the

graph. These two modes may correspond to different sorts of behavior by the Chair in his/her interactions with panel members during the meeting.

Among the many roles of the committee chair is to serve as a facilitator, ensuring that all of the panel members present are able to express their views. The chair's role in "flattening" the structure of a given meeting's graph could suggest a particular facilitation strategy, wherein the committee chair tries to elicit the opinions of voting members who speak later, and might otherwise be less influential. When the chair does not act to significantly change the graph structure, the chair may be taking on the role of a synthesizer – gathering the opinions of the other voters to answer FDA questions, but not causing any of the other members to repeat what has already been said. The histogram shown in Figure 66, suggests that there may be two different strategies that committee chairs could use during a meeting.

24. Committee chair impact is significantly positively associated with meeting date for meetings in which there is a voting minority.

The marked change in committee chair strategy has many potential explanations – one might be that the identity of the committee chair changed around this time, suggesting a change in personal style, but this would have to explain a consistent change across multiple chairs. Alternatively, prior to 2002, most of the meetings were chaired by women, whereas after March 2002 most chairs were men. There is literature to suggest that men and women utilize different leadership styles on committees (e.g., Johnson & Eagley 1990). Another explanation might be that there was a change in FDA policy regarding committee operations. For example, there was a change in conflict of interest reporting procedures that was concurrent with the shift shown in empirical result 22, but there is no obvious connection between these two events. Another possibility is that the types of

devices that came to the Circulatory Systems Devices Panel changed around this time – this could be reflected in the fact that there were no half-day meetings after 2002. Perhaps there was an increase in the difficulty of the devices that the panel reviewed (e.g., concurrent with the entrance on the market of drug-eluting stents, Left Ventricular Assist Devices and other potentially risky devices). Finally, we might hypothesize some combination of these ideas – that a change in FDA policy might have some way impacted upon chair style, and that this change in policy might have been driven by changing market conditions. Testing these hypotheses requires looking across multiple panels and committees, which we leave to future work.

Conflicts of Interest

25. Panel members with conflicts of interest tend to have a significantly higher h-index than do panel members without a conflict.

This finding recognizes that panel members with conflicts of interest tend to have greater academic credentials than do other panel members. When a particular panel member is granted a waiver by the FDA despite their conflict of interest, the FDA recognizes the impossibility of entirely eliminating conflicts of interest on panels of experts. The need for specialized expertise often requires that individuals who have extensive experience with a device be consulted. On the other hand, it is precisely these individuals who are likely to have financial conflicts due to their previous work. (McComas et al. 2005) call this the "shared-pool dilemma".

26. Panel members with conflicts of interest do not have higher outdegrees than panel members without conflicts of interest.

27. Panel members with conflicts of interest, who are in the voting majority do not have higher outdegrees than panel members with conflicts of interest who are in the voting minority.

Procedures on the Circulatory Systems Devices Panel seem to be unaffected by the conflicts of interest outlined above. Indeed, there seems to be no evidence of systemic bias due to conflict of interest on this panel. Further investigation therefore focuses on the level of individual meetings. The evidence here suggests that conflicts of interest are minimized when there is only one member on the panel with a reported conflict or when there are opposing reported conflicts on the panel. When there are multiple panel members with consistent conflicts of interest, unanimous support for those conflicts might result. Although more data is required to rigorously test this finding, if this speculation is correct, it suggests a direction for future research on conflicts of interest. In particular, trends in the data seem to indicate that, where there is one panel member with an identified conflict, or when there are opposing conflicts on a panel, these conflicts are typically neutralized. This is reflected in the first part of empirical findings 26. One possible explanation for this phenomenon is that the prospect of appearing biased might cause panel members with conflicts of interest to become more aware of how their conflicts might drive their decision-processes. On the other hand, when consistent conflicts are distributed across many panel members, it is less likely that any one individual will question another's conflict. Another explanation could be that a given device is unambiguously approvable or non-approvable – under these circumstances, the conflicts of interest might simply be consistent with the device's qualities. Future work will focus on testing these hypotheses further.

Synthesis of Empirical Results

Because committee members have scarce attention resources (Cobb and Elder 1982), procedure is necessary to ensure that the appropriate members are given ample time to speak. A sociologist would characterize these procedures as embodying a status hierarchy – nevertheless, we do not use the standard status assumptions that are found in the sociology literature. This is because empirical results do not agree with the hypothesis that some of these variables (e.g., race and gender) significantly impact of air-time in the direction predicted (i.e., women tend to speak more than men). Festinger's theory of group decision-making (1954) suggests that these variables are not the source of significant differences because they are not perceived as directly relevant to a technical query. Panel procedures likely act to focus attention of panel members to the question at hand, potentially explaining the effects of h-Index, a metric reflecting recognized academic expertise. Panel procedures may represent a form of common knowledge (cf. Chwe 2003) that may encode assumptions regarding perceived expertise. Thus, procedural choices may modulate the flow of communication on a committee in ways that can enable or disable the covariance of actual influence and perceived expertise.

A generalized relation between panel procedure and actual influence could be supported by the empirical observation that members of the voting minority tend to have a lower graph outdegree than do members of the voting majority. A covariance between actual influence and perceived expertise could be reflected in the fact that voting minority members also tend to be positioned later in the speaking order than do their voting majority counterparts. Nevertheless, this approach might lead to a situation in which some experts are marginalized – e.g., if the majority of committee members are unwilling to listen to an opinion that might have some merit but disagrees with their intuition. This effect would reduce some of the benefit from diversity shown by Hong and Page (2004). In

such circumstances, a respected mediator, such as the committee chair, might take action to promote the exchange and re-consideration of information that might otherwise be ignored by the majority. That such chair behavior indeed occurs on FDA panels is reflected in the "flattening" of the panel hierarchy as seen in our directed graphs by the committee chair.

The above assumes that an expert committee member is able to determine whether another member is correct in his or her comments. This assumption may break down in the case of very different technical specialties, wherein a member from one specialty has limits in evaluating the contribution of a member from another specialty. This reflects a sort of cognitive limitation that can prevent a perfect flow of information, even in the face of clear procedure (Douglas 1986). This effect is reflected in the empirical observation that members of different medical specialties have different speaking habits. Furthermore, members of a given medical specialty seem to preferentially link to one another on FDA panels. Nevertheless, this effect is not incommensurable, as indicated by the fact that the empirical distribution tends to place more probability mass near 0.5 than does the simulated distribution. These results suggest that medical specialty plays a role as an informal constraint on communication (North 2006/1990), which may be circumvented by various mechanisms including cross-specialty understanding, or if another identity is evoked. On one hand, evocation of a common identity might further enable learning among panel members. On the other hand, as individuals' language deviates from their assigned roles, panel members may not know how to evaluate the contributions of others. Furthermore, in the most uncertain meetings, in which there is a minority of size two or more, we find that when vote cohesion among specialists of a given type is high so is specialty cohesion. This suggests that, under conditions of high uncertainty, individuals might engage in role-based behavior (March 1994), relying on their medical specialty identification to determine what to talk about and how to vote. On the

other hand, if this identification is not found to be relevant (i.e., there is a low specialty cohesion percentile) vote cohesion percentile is also low. This suggests circumstances in which panel members' voting behavior is not related to their expressed concerns, potentially indicating voting for idiosyncratic reasons. That voting cohesion percentile is low when specialty cohesion is low suggests that medical specialty is an organizing factor on these panels. For example, if one were to think of the data regarding a particular medical device as a correlation device (e.g., a coin flip, cf. Aumann 1974) indicating data quality, then an individual's subjective interpretation of that data would be conditional on medical specialty. The strength of this conditional dependence is a decreasing function of ambiguity.

Empirical results demonstrate that meeting length, minority size, minority outdegree, and propensity for the lead reviewers to disagree are correlated. This suggests the presence of an underlying explanation for this effect. We propose that these are all related to the difficulty of evaluation of a given device. For example, the data might be ambiguous or it might not be clear what constitutes a correct decision. This is a form of deep uncertainty that panels such as these occasionally face – the information necessary to make a clear choice may simply not be available because the right expertise may not be possessed by panel members, or because it may simply not exist. In such cases, committee members are likely to be unsure which role is appropriate, and may instead rely upon idiosyncratic opinions and values (March 1994). Meeting length would increase as panel members discuss different perspectives, whereas the added controversy could also result in larger and more influential minorities (i.e., more dissent).

Modeling Results and Their Implications

In Chapter 6, we presented a computational model that attempts to capture some of the dynamics described above. Although still preliminary, the model yields

nine results that may provide some theoretical insight into the empirical trends discussed above.

1. Simulated vote cohesion percentile and simulated specialty cohesion percentile are significantly positively associated.

   As in the empirical data, this result suggests that medical specialty is an organizing factor in determine voting behavior on panels in which there is no consensus. The model posits that panel members from the same medical specialty are more likely to understand information that is consistent with their training, thereby shaping their voting behavior.

2. Proportional minority size, and specialty and vote cohesion, are functions of device quality.

This demonstrates that proportional minority size is associated with device quality; an intrinsic characteristic of the device being reviewed.

3. Members of the simulated voting minority tend to speak later than do members of the simulated voting majority when only when speaking order is pre-set. Furthermore, simulated vote and specialty cohesion percentile distributions fit the data better when speaking order is pre-set. Correlation between specialty cohesion percentile and vote cohesion percentile is slightly stronger than under conditions of pre-set speaking order.

These results suggest one possible mechanism by which the empirical speaking order effect on vote might be explained. In particular, speaking order and seating order may be chosen in advance as one means by which the FDA could exercise control over the process. Reasons why this might be the case were presented in the discussion of the empirical results. If such is the case, then speaking order

might represent a coordination mechanism by which FDA expresses common knowledge to panel members (Chwe 2003). Of course, other explanations that are not captured in the model are possible. Future work may focus on determining a mechanism of emergence of voting effects on speaking order.

4. The model better explains the empirical data distributions when no advocacy among panel members is assumed, although correlation between specialty cohesion percentile and vote cohesion percentile is stronger than under conditions of non-advocacy.

These results suggest that the model better represents reality under conditions of non-advocacy. This suggests the possibility that discussion on such panels is indeed motivated by learning or information sharing rather than strategic concerns. That specialty and voting cohesion are more strongly correlated under conditions of advocacy is not surprising, since panel members are less likely to share information. When specialties are relatively homogeneous in their preferences, this would reduce the possibility that information is shared across specialty boundaries, perhaps leading to a "group think" within each specialty. On the other hand, when specialties' preferences are internally diverse, specialty groups would fragment since information sharing would be limited. Furthermore, links would be unlikely to form across specialty boundaries between members who vote the same way because simulated panel members with broad expertise might be unwilling to discuss relevant information.

5. Simulated complexity, mean breadth, mean depth, depth dispersion, openness and ambiguity all have a significant effect on simulated specialty cohesion percentile.

These simulated results suggest potential drivers of the results seen in the empirical data. In particular, we find that specialty cohesion percentile decreases

with mean breadth and openness, and increases with complexity, mean depth and depth dispersion. We may explain these findings by noting that panel members who share deep domain expertise will be more likely to discuss that shared knowledge with members of the same specialty. Members of other specialties may not be able to learn this specialized information and so will not be linked. This increases the propensity for panel members from the same specialty to link in a conversation while decreasing the propensity for links across specialties. A similar argument holds for depth dispersion. A high dispersion in depth increases the likelihood that at least one panel member will have a very high depth, and then share the resulting knowledge with other members of that specialty. Panel members have more topics to discuss with increasing complexity, and therefore are less likely to discuss all topics, potentially leading to a preferential discussion of those topics in their specialty – especially since they are more readily able to assimilate these topics from others, potentially creating a "hidden profile" effect of the sort described by Stasser and Titus (1985). They are also likely to speak longer, creating more of an opportunity for members from the same specialty to link. A similar argument holds for ambiguity – as ambiguity increases, panel members spend more time in discussion, enabling linkage between members of the same specialty. For very low values of ambiguity, panel members have relatively little to discuss and so do not require linkage within their specialty groups. Unambiguous device data leads to an early consensus. Furthermore, discussion time is curtailed. On the other hand, as mean breadth increases, panel members are more likely to link across specialty boundaries, reducing specialty cohesion. Finally, as openness increases, panel members are less constrained to follow a fixed speaking order. This reduces the degree to which initial speakers can cause later speakers to adopt favored topics. As a result, specialties may be less aligned since they will be less likely to share, and therefore discuss, common information in their specialties.

6. Simulated complexity and ambiguity have a significant effect on simulated voting cohesion percentile.

These simulated results suggest that as complexity and ambiguity increase, vote cohesion increases. This is consistent with the previous argument because vote cohesion is only defined for those meetings in which there is a sizable voting minority. In these meetings, it is likely that discussion causes panel members who exchange information to align their preferences. If there is not enough time for this to occur, voting cohesion will be low. As ambiguity and complexity increase, there is more information shared, since panel members are more able to discuss their specific unshared knowledge. This could lead to more learning, and ultimately, higher vote cohesion. Excluded are the cases in which the panel reaches consensus. These are typically associated with low ambiguity. Thus the association of high vote cohesion with high ambiguity and complexity suggests the formation of coherent subgroups on the panel that are unable to reach consensus. This could indicate the presence of multiple competing interpretations of the data.

7. Association between specialty cohesion percentile and vote cohesion percentile is stronger when ambiguity is less than 0.5 and when complexity is less than 20 topics.

This result suggests that in the presence of high complexity and ambiguity, the correlation between medical specialty and vote cohesion percentiles breaks down. Although specialty cohesion percentile and vote cohesion percentile are both positively associated with ambiguity and complexity, their association with each other decouples. One possible explanation is that, under these conditions, there are too many topics to discuss as well as conflicting signals from the data. Even though there is more time to discuss these issues, it is unlikely that panel members will cover enough ground to reach a consensus within their specialty.

This might suggest a high specialty cohesion – i.e., among members in a specialty that share similar topics – but a low vote cohesion, since there may be subgroups within specialties. If there is a breadth of expertise on the panel, learning across specialties would strengthen this trend, because the learning would be selective. Furthermore, the presence of conflicting signals suggests that a small difference in valence in one topic could a panel member's vote. This would suggest high specialty cohesion but low vote cohesion.

8. Simulated proportional minority size is significantly associated with diversity, complexity, mean breadth, quality and ambiguity.

In particular, proportional minority size increases with diversity and ambiguity, and decreases with complexity, mean breadth, and quality. This makes sense because, as diversity increases, there are more barriers to communication across specialty boundaries. This is either because there are more specialties which may not share knowledge with one another *a priori*. Furthermore, as ambiguity increases, different topics give different signals. Thus voting behavior is highly sensitive to the information accessed by a particular voting member. As complexity increases, there is more time for panel members to discuss different topics, leading to a higher likelihood of information sharing. Furthermore, there are more topics to discuss. As a result, it is more likely that a speaker will discuss a previously unmentioned topic, leading to a situation in which panel members could learn. As mean breadth increases, information is more likely to be communicated across specialty boundaries, leading to a common understanding of the data, and consequently, a higher probability of consensus. Finally, as quality moves away from zero, topics are more likely to display similar signals, leading to less reason for disagreement *a priori*.

9. Simulated meeting profile, complexity, mean depth, depth dispersion, openness, ambiguity and quality are all significantly associated with the panel's capacity to reach a correct decision.

In particular, as mean depth and depth dispersion increase, the available information also increases, and that information becomes more accurate and more likely to be correctly transmitted. Furthermore, as the absolute value of quality increases, and as ambiguity decreases, different topics display consistent, correct signals to all panel members, leading to a consensus on the correct outcome. As complexity increases, panel members are more likely to spend time in discussion, and therefore more likely to come to a correct conclusion. It is likely that very low values of complexity lead to short discussion times, suggesting that a panel's decision might not be well-informed. For example, one panel member might have access to a depth of unshared expertise that other panel members ignore. This is interesting because it indicates that many "easy" decisions might require deeper consideration than they are given. Often, prolonged discussion may raise issues that might have otherwise gone unnoticed. A similar argument holds for the impact of openness. If openness is too high, then panel members will be unable to make extended arguments in a given topic. In the model, this manifests as repeated mention of a topic, increasing its probability of adoption. On the other hand, with a well-established structure, panel members have sufficient time to ensure that the full range of their topics is heard. Otherwise, the panel meeting might end early, with the panel having achieved total- or near-consensus with an uninformed perspective.

These initial results serve as a benchmarking for the model that may be expanded upon in future work. The fit of the model to some aspects of the data, even using randomized parameters, suggests that some major elements on FDA panels are captured. These include the following:

1) The "technical" effects of ambiguity and complexity – It is interesting that the correlation between specialty cohesion percentile and vote cohesion percentile erodes under conditions of high ambiguity and complexity, as panel members speak about very different topics and draw different conclusions from the same topic. This is a finding that is generally consistent with the theories of March (1994). This suggests that a limit placed upon interpretation of empirical reality erodes as that reality becomes harder to discern.

2) The "cognitive" effects of learning and medical specialty – The model is consistent with the notion that medical specialty is an organizing factor. Vote and specialty cohesion are correlated following model rules. The model assumes that communication between medical specialties can only occur via the mechanism of individuals who possess a breadth of expertise – i.e., a capacity to learn from and teach other in different medical specialties. The role of mean breadth in decreasing proportional minority size and specialty cohesion percentile points to another sort of learning on panels that, when combined with high depth, could enable better information aggregation across specialty boundaries. That mean breadth is not associated with voting cohesion percentile is a reflection of the fact that voting cohesion percentile may not exist in these cases – the panel may instead have reached consensus. The absence of an effect of breadth on correct voting outcome suggests that learning requires a depth of expertise on which to draw in order for information to be successfully transferred – absent the appropriate depth of expertise, breadth of expertise is not useful. Indeed, proximity to ground truth on a topic is a function of breadth and not depth. If there is no depth of expertise on the panel, breadth will not lead to the correct answer.

3) The "social" effects of speaking order and procedure – It is interesting that the model only matches the data when speaking order is pre-determined, such that members in the voting minority already speak last. This provides one possible explanation for the observed behavior in the FDA panels. Nevertheless, other explanations, that do not assume an explicit pre-set ordering, are possible (see, e.g., Chwe 2000). It is not inconceivable that this result could instead have emerged due to another mechanism such as a tendency for late speakers to want to disagree e.g., due to resentment over being assigned a late position. Furthermore, a structured panel process with sufficient discussion time may be important in enabling panel members to express information that would otherwise be unshared. On the other hand, if much of the information discussed in the structured phase is already known by other panel members, or if ambiguity is very low, then a strict structure to the discussion might not be beneficial since there is little that panel members could learn. In such a case, open discussion would better serve the goals of the panel conveners by avoiding the repetition of shared information (potentially leading to a quick, but potentially uninformed, decision in favor of what this information represents).

We may draw some concluding observations from the analysis that we have already performed. If the logic underlying our model is correct, the Circulatory Systems Devices Panel in the FDA may largely allow for learning and correct decision-making. Empirical results suggest that three major factors affect panel decision-making; namely, technology, procedure and training. In general, these are representative of three interacting layers, or orders (cf. Hayek 1952, Richards 2009, see Figure 76): respectively, the technical, cognitive, and social. Within the technical layer, data are important – this is where data ambiguity and device quality can impact upon decision outcomes. Within the cognitive layer, direction

of attention is important – this is where learning may occur as modulated by the effects of medical specialty. Finally, within the social layer, dynamics of perceived expertise are strong – this is where speaking order becomes a determinant of voting behavior. No one of these layers on its own is sufficient to explain voting behavior. Instead, interactions among these layers cause complex social behavior of the sort required to successfully cope with a complex technical environment (Conway 1968).

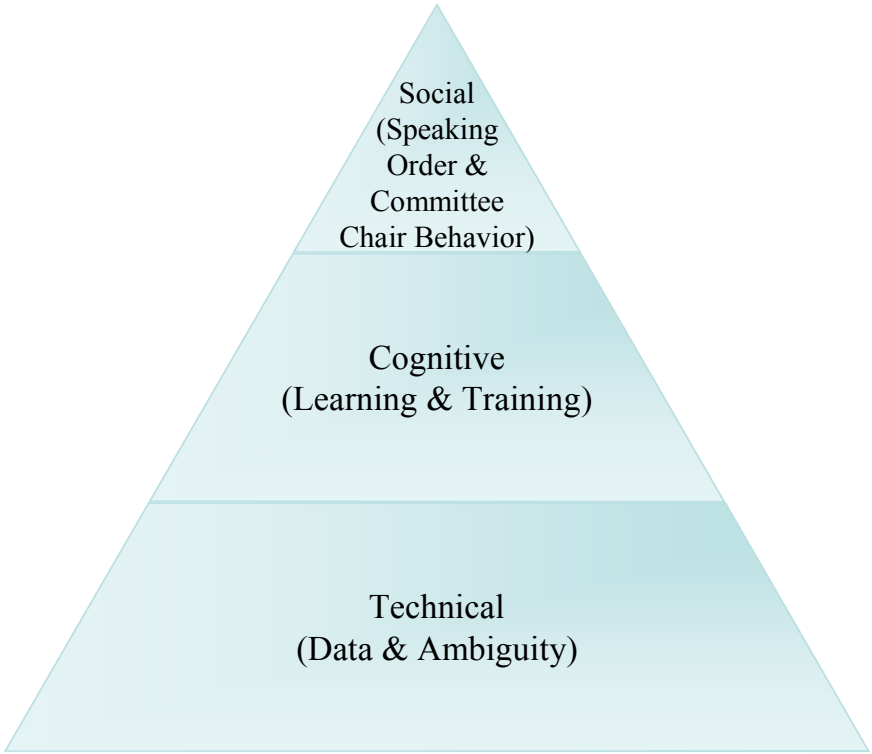


Figure 76: Three interacting orders or layers. Each one constrains the layer immediately above it (cf. Polanyi 1970). Thus very clear data would constrain the set of possible interpretations, etc.

Today's engineering systems must be able to adapt quickly to an increasingly complex world. Only by pooling knowledge from across many different domains can this be accomplished. Still, there has been little empirical research into how this occurs in real-world settings. We addressed this query by asking three questions that guide our research:

1. How can we study, in a quantitative, consistent manner, the flow of communication among technical experts on committee decisions?

2. How do technical experts' decisions change as they learn and interact during the decision-making process?

3. How might we design committee processes so as to enable desirable behaviour on the part of technical expert committees?

The methodological contribution of this thesis answers question 1, providing a tool that may be used by future researchers to study verbal communication flows on expert committees. This tool is used to inform theory, which has been instantiated as a preliminary computational model, thus providing a first answer to question 2. As more data is gathered and analyzed, we will be even more able to answer question 2, providing insight into the decision- and learning-processes of technical experts. Finally, this chapter provides a preliminary framework for question 3. Together, these diverse sources of information can be combined to lead us to a deeper understanding of committee decision-making on technical expert committees, and ultimately, to the better design of engineering systems.

BIBLIOGRAPHY

Policy and Guidance Handbook for FDA Advisory Committees. (1994). National Technical Information Service of the U.S. Department of Commerce (pp. 1-418). Rockville, MD: U.S. Food and Drug Administration.

Arrow, K. (1963). *Social Choice and Individual Values* (Vol. 2).

Ashworth, S., & Bueno de Mesquita, E. (2005). Valence Competition and Platform Divergence. *Princeton University Typescript*.

Aumann, R. J., & Dreze, J. H. (1974). Cooperative games with coalition structures. *International Journal of Game Theory*, *3*(4), 217-237. doi: 10.1007/BF01766876.

Aumann, R. J. 1974. Subjectivity and correlation in randomized strategies. *Journal of Mathematical Economics* 1, no. 1. Journal of Mathematical Economics: 67-96.

Axelrod, R. M. (1976). *Structure of Decision: The Cognitive Maps of Political Elites*. Princeton Univ Pr.

Báles, R. F., Strodtbeck, F. L., Mills, T. M., & Roseborough, M. E. (1951). Channels of Communication in Small Groups. *American Sociological Review*, *16*(4), 461-468. doi: 10.2307/2088276.

Baron, D. P., & Ferejohn, J. A. (1989). Bargaining in Legislatures. *The American Political Science Review*, *83*(4), 1181-1206.

Benjamin, W. (1969). *Illuminations: Essays and Reflections* (first Schocken paperback edition.). Schocken.

Bentham, J. (1988/1780). *The Principles of Morals and Legislation*. Amherst, New York: Prometheus Books.

Berger, J., Cohen, B. P., & Zelditch, M. (1972). Status Characteristics and Social Interaction. *American Sociological Review*, *37*(3), 241-255. doi: 10.2307/2093465.

Black, D. (1948). On the Rationale of Group Decision-making. *The Journal of Political Economy*, *56*(1), 23-34.

Blei, D. M., & Lafferty, J. D. (2006). Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning* (pp. 113-120). Pittsburgh, Pennsylvania: ACM. doi: 10.1145/1143844.1143859.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*, 993-1022.

Boffey, P. (1976). Scientists and Bureaucrats: A Clash of Cultures on FDA Advisory Panel. *Science*, *191*(4233), 1244-1246.

Boroditsky, L. (2002). You are what you speak. *NewScientist*, 34-38.

Boroditsky, L. (2003). Linguistic Relativity. In *Encyclopedia of Cognitive Science* (pp. 917-922). London: Macmillan.

Bottger, P. C. (1984). Expertise and air time as bases of actual and perceived influence in problem-

solving groups. *Journal of Applied Psychology*, *69*(2), 214-221.
Broniatowski, D. A., & Weigel, A. L. (2006). *Decision-Making in the Political and Technical Environments*. Cambridge, MA: Massachusetts Institute of Technology.
Brown, J. A. (1986). Professional language: words that succeed. *Radical History Review*, *34*, 33-51.
Carley, K. M. (1997). Extracting team mental models through textual analysis. *Journal of Organizational Behavior*, *18*, 533–558.
Carley, K. M. (2003). Dynamic network analysis. In *Dynamic social network modeling and analysis: Workshop summary and papers* (pp. 133–145).
Carley, K., & Palmquist, M. (1992). Extracting, representing, and analyzing mental models. *Social Forces*, *70*(3), 601–636.
Carlile, P. R., & Schoonhoven, C. B. (2002). A Pragmatic View of Knowledge and Boundaries: Boundary Objects in New Product Development. *Organization Science*, *13*(4), 442-455.
Chwe, M. S. (2000). Communication and Coordination in Social Networks. *The Review of Economic Studies*, *67*(1), 1-16.
Chwe, M. S. (2003). *Rational Ritual: Culture, Coordination, and Common Knowledge*. Princeton University Press.
Cobb, R. W., & Elder, C. D. (1983). *Participation in American Politics: The Dynamics of Agenda-Building* (p. 196). Baltimore and London: The Johns Hopkins University Press.
Cohn, C. (1987). Sex and Death in the Rational World of Defense Intellectuals. *Signs*, *12*(4), 718, 687
Conway, M. E. (1968). How Do Committees Invent? *Datamation*.
Corman, S. R., Kuhn, T., Mcphee, R. D., & Dooley, K. J. (2002). Studying Complex Discursive Systems. *Human Communication Research*, *28*(2), 157-206. doi: 10.1111/j.1468-2958.2002.tb00802.x.
Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*, *41*(6), 391-407.
Dhillon, I. S., & Modha, D. S. (2001). Concept Decompositions for Large Sparse Text Data Using Clustering. *Machine Learning*, *42*, 143-175.
Diesner, J., & Carley, K. M. (2004). *AutoMap1. 2-Extract, analyze, represent, and compare mental models from texts*. Carnegie Mellon University, School of Computer Science, Institute for Software Research International, Technical Report CMU-ISRI-04-100.
Dong, A. (2005). The latent semantic approach to studying design team communication. *Design Studies*, *26*, 445-461.
Dong, A., Hill, A. W., & Agogino, A. M. (2004). A Document Analysis Method for Characterizing Design Team Performance.

Journal of Mechanical Design, *126*, 378-385.
Douglas, M. (1986). *How Institutions Think*. Syracuse, New York: Syracuse University Press.
Douglas, M., & Wildavsky, A. (1982). *Risk and Culture* (p. 221). Berkeley, CA: University of California Press.
Dumais, S. T. (1991). Improving the retrieval of information from external sources. *Behavior Research Methods, Instruments, & Computers*, *23*(2), 229-236.
Dupret, G. (2003). Latent concepts and the number orthogonal factors in latent semantic analysis (pp. 221 - 226). Toronto, Canada : ACM.
Eco, Umberto. 1997. *The Search for the Perfect Language*. Wiley-Blackwell, April 15.
Eco, U., & McEwen, A. (2001). *Experiences in translation*. University of Toronto Press.
Elder, C. D., & Cobb, R. W. (1983). *The Political Uses of Symbols*. Longman professional studies in political communication and policy. New York: Longman.
Fader, A., Radev, D., Crespin, M. H., Monroe, B. L., Quinn, K. M., & Colresi, M. (2007). MavenRank: Identifying Influential Members of the US Senate Using Lexical Centrality (pp. 658-666). Prague: Association for Computational Linguistics.
Fararo, T. J., & Skvoretz, J. (1986). E-State Structuralism: A Theoretical Method. *American Sociological Review*, *51*(5), 591-602.
Festinger, L. (1954). A theory of social comparison processes. *Human relations*, *7*(2), 117-140.
de Finetti, B. (1974). Theory of probability. Vol. 1-2. *English translation, Wiley*.
Fişek, M. H., Berger, J., & Norman, R. Z. (1991). Participation in Heterogeneous and Homogeneous Groups: A Theoretical Integration. *The American Journal of Sociology*, *97*(1), 114-142. doi: 10.2307/2781640.
Foltz, P. W., Kintsch, W., & Landauer, T. K. (1998). The Measurement of Textual Coherence with Latent Semantic Analysis. *Discourse Processes*, *25*(2&3), 285-307.
Frey, D., Herder, P., Wijnia, Y., Subrahmanian, E., Katsikopoulos, K., & Clausing, D. (2009). The Pugh Controlled Convergence method: model-based evaluation and implications for design theory. *Research in Engineering Design*, *20*(1), 41-58. doi: 10.1007/s00163-008-0056-z.
Friedman, R. S. (1978). Representation in Regulatory Decision Making: Scientific, Industrial, and Consumer Inputs to the F.D.A. *Public Administration Review*, *38*(3), 205-214.
Gaertner, W. (2009). *A Primer in Social Choice Theory* (Revised edition.). Oxford University Press.
Galbraith, J. R. (1993). *Competing with Flexible Lateral Organizations* (2nd ed.). Prentice Hall.
Gansner, E. R., & North, S. C. (1999). An Open Graph Visualization

System and Its Applications to Software Engineering. *SOFTWARE - PRACTICE AND EXPERIENCE*, *30*, 1203--1233.

Garfinkel, H. (1984). *Studies in ethnomethodology*. Wiley-Blackwell.

Gelijns, A. C., Brown, L. D., Magnell, C., Ronchi, E., & Moskowitz, A. J. (2005). Evidence, Politics, And Technological Change. *Health Affairs*, *24*(1), 29-40.

Gibson, D. R. (2003). Participation Shifts: Order and Differentiation in Group Conversation. *Social Forces*, *81*(4), 1335-1380. doi: 10.2307/3598117.

Gibson, D. R. (2005). Taking Turns and Talking Ties: Networks and Conversational Interaction. *American Journal of Sociology*, *110*(6), 1561-1597. doi: 10.1086/428689.

Gibson, D. R. (2008). How the Outside Gets in: Modeling Conversational Permeation. *Annual Review of Sociology*, *4*.

Goffman, E. (1981). *Forms of Talk*. Blackwell Publishers.

Goodman, N., Mansinghka, V., Roy, D. M., Bonawitz, K., & Tenenbaum, J. (2008). Church: a language for generative models with non-parametric memoization and approximate inference. In *Uncertainty in Artificial Intelligence*.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(Suppl 1), 5228-5235. doi: 10.1073/pnas.0307752101.

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in Semantic Representation. *PSYCHOLOGICAL REVIEW-NEW YORK-*, *114*(2), 211.

Hall, D., Jurafsky, D., & Manning, C. D. (2008). Studying the History of Ideas Using Topic Models. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing* (pp. 363-371). Association for Computational Linguistics.

Hare, A. P., & Bales, R. F. (1963). Seating position and small group interaction. *Sociometry*, *26*(4), 480-486.

Hayek, F. A. 1952. *The sensory order*. University of Chicago Press Chicago.

von Herder, J. G. (2002/1767). *Herder: Philosophical Writings* (1st ed.). Cambridge University Press.

Hill, A. W., Dong, A., & Agogino, A. M. (2002). Towards Computational Tools For Supporting the Reflective Team (pp. 305-325). Dordrecht, Netherlands: Kluwer.

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(46), 16569-16572. doi: 10.1073/pnas.0507655102.

Hofmann, T. (2001). Unsupervised Learning by Probabilistic Latent Semantic Analysis. *Machine Learning*, *42*, 177-196.

Hong, L., & Page, S. E. (2004). Groups of diverse problem

solvers can outperform groups of high-ability problem solvers. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(46), 16385-16389. doi: 10.1073/pnas.0403723101.

von Humboldt, W. F (1997/1820). *Essays on Language*. Peter Lang Publishing.

Jasanoff, S. S. (1987). Contested Boundaries in Policy-Relevant Science. *Social Studies of Science*, *17*(2), 195-230.

Johnson, B., & Eagly, A. (1990). Gender and Leadership Style: A Meta-Analysis. *CHIP Documents*. Retrieved April 7, 2010, from http://digitalcommons.uconn.edu/chip_docs/11.

Kahneman, D., & Tversky, A. (1979). Prospect Theory: An Analysis of Decision Under Risk. *Econometrica*, *47*(2), 263-292.

Kameda, T., Ohtsubo, Y., & Takezawa, M. (1997). Centrality in Sociocognitive Networks and Social Influence: An Illustration in a Group Decision-Making Context. *Journal of Personality and Social Psychology*, *73*(2), 309, 296.

Kaptchuk, T. J. (2003). *Effect of interpretive bias on research evidence* (Vol. 326, pp. 1453-1455). BMJ Publishing Group Ltd.

Kelly, C. D., & Jennions, M. D. (2006). The h index and career assessment by numbers. *Trends in Ecology & Evolution*, *21*(4), 167-170. doi: 10.1016/j.tree.2006.01.005.

Landauer, T. K., & Dumais, S. T. (1997). A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction and Representation of Knowledge. *Psychological Review*, *104*(2), 211-240.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An Introduction to Latent Semantic Analysis. *Discourse Processes*, *25*, 259-284.

Landauer, T. K., Laham, D., & Derr, M. (2004). From paragraph to graph: Latent semantic analysis for information visualization. *Proceedings of the National Academy of Science*, *101*(1), 5214-5219.

Latane, B. (1981). The psychology of social impact. *American Psychologist*, *36*(4), 343–356.

Latane, B. (1996). Dynamic social impact: The creation of culture by communication. *Journal of Communication*, *46*, 13–25.

Lawson, C.M. (2008) Group decision making in a prototype engineering system : the Federal Open Market Committee. (2008, June). . Thesis and Exposé of the Golden Path Method. Retrieved January 28, 2009, from http://dspace.mit.edu/handle/1721.1/43854.

Lunney, G. H. (1970). Using Analysis of Variance with a Dichotomous Dependent Variable: An Empirical Study. *Journal of Educational Measurement*, *7*(4), 263-269.

Luo, J., Whitney, D., Baldwin, C. Y., & Magee, C. L. (2009). Measuring and Understanding Hierarchy as an Architectural Element in

Industry Sectors. *SSRN eLibrary*. Retrieved February 9, 2010, from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1421439.

Lurie, P., Almeida, C. M., Stine, N., Stine, A. R., & Wolfe, S. M. (2006). Financial Conflict of Interest Disclosure and Voting Patterns at Food and Drug Administration Drug Advisory Committee Meetings. *Journal of the American Medical Association*, *295*(16), 1921-1928.

Maisel, W. H. (2004). Medical Device Regulation: An Introduction for the Practicing Physician. *Annals of Internal Medicine*, *140*(4), 296-302.

Maisel, W. H. (2005). A Device for Proximal Anastomosis of Autologous Coronary Vein Grafts: Report from the Meting of the Circulatory System Devices Panel of the Food and Drug Administration Center for Devices and Radiologic Health. *Circulation*, (112), 1516-1518.

Manning, C. D., & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing* (1st ed.). The MIT Press.

March, J. G. (1994). *A Primer on Decision Making: How Decisions Happen* (p. 289). New York, NY: The Free Press.

Maynard, D. W. (1980). Placement of topic changes in conversation. *Semiotica La Haye*, *30*(3-4), 263-290.

McCallum, A., Wang, X., & Corrada-Emmanuel, A. (2007). Topic and role discovery in social networks with experiments on Enron and academic email. *Journal of Artificial Intelligence Research*, *30*, 249-272.

McComas, K. A., Tuite, L. S., & Sherman, L. A. (2005). Conflicted scientists: the "shared pool" dilemma of scientific advisory committees. *Public Understanding of Science*, *14*, 285-303.

McMullin, H., & Whitford, A. B. (2007). Extra-Judicial Decision Making for Drug Safety and Risk Management: Evidence from the FDA. *Northwestern Journal of Technology and Intellectual Property*, *5*(2), 250-264.

Mulkay, M., Pinch, T., & Ashmore, M. (1987). Colonizing the Mind: Dilemmas in the Application of Social Science. *Social Studies of Science*, *17*(2), 231.

Nelson, R. R. (2005). On the Uneven Evolution of Human Know-how. In *Technology, Institutions, and Economic Growth* (pp. 173-194). Cambridge, Massachusetts: Harvard University Press.

Neuman, W. L. (2005). *Social Research Methods: Quantitative and Qualitative Approaches* (6th ed.). Allyn & Bacon.

von Neumann, J., & Morgenstern, O. (2007). *Theory of Games and Economic Behavior (Commemorative Edition)* (60th ed.). Princeton University Press.

Newman, D. J., & Block, S. (2006). Probabilistic Topic Decomposition of an Eighteenth-Century American Newspaper. *Journal of the American Society for Information Science and Technology*, *57*(6), 753-767.

Nonaka, I., Toyama, R., & Konno, N. (2000). SECI, Ba and Leadership: a Unified Model of Dynamic Knowledge Creation. *Long Range Planning*, *33*(1), 5-34. doi: 10.1016/S0024-6301(99)00115-6.

North, D. C. (1990). *Institutions, Institutional Change and Economic Performance*. Cambridge University Press.

Nowak, A., Szamrej, J., & Latane, B. (1990). From private attitude to public opinion: A dynamic theory of social impact. *Psychological Review*, *97*(3), 362–376.

Okamoto, D. G., & Smith-Lovin, L. (2001). Changing the Subject: Gender, Status, and the Dynamics of Topic Change. *American Sociological Review*, *66*(6), 852-873. doi: 10.2307/3088876.

Page, S. E. (2008). *The Difference: How the Power of Diversity Creates Better Groups, Firms, Schools, and Societies* (1st ed.). Princeton University Press.

Papadimitriou, C. H., Raghavan, P., Tamaki, H., & Vempala, S. (2000). Latent Semantic Indexing: A Probabilistic Analysis. *Journal of Computer and System Sciences*, (61), 217-235.

Parisian, S. (2001). *FDA Inside and Out* (p. 647). Front Royal, VA: Fast Horse Press.

Peirce, C. S. (1934-48). *Collected* Papers (4 vols.). Cambridge: Harvard University Press.

Pentland, A. S. (2008). *Honest signals: how they shape our world*. The MIT Press.

Pines, W. L. (2002). *How to Work With the FDA: Tips from the Experts* (p. 241). Washington, DC: FDLI.

Polanyi, M. (1958). *Personal Knowledge: Towards a Post-Critical Philosophy* (p. 428). Chicago, IL: University of Chicago Press.

Polanyi, M. (1970). Transcendence And Self-Transcendence. *Soundings* 53, no. 1 (Spring): 88-94.

Pole, A., West, M., & Harrison, J. (1994). *Applied Bayesian Forecasting and Time Series Analysis* (1st ed.). Chapman and Hall/CRC.

Quinn, K. M., Monroe, B. L., Colaresi, M., Crespin, M. H., & Radev, D. R. (2006). An Automated Method of Topic-Coding Legislative Speech Over Time with Application to the 105th-108th US Senate.

Rapp, R. (2000). Extra chromosomes and blue tulips: medico-familial interpretations. *Living and Working with the New Medical Technologies: Intersections of Inquiry*.

Rehder, B., Schreiner, M. E., Wolfe, M. B. W., Laham, D., Landauer, T. K., & Kintsch, W. (1998). Using Latent Semantic Analysis to assess knowledge: Some technical considerations. *Discourse Processes*, *25*(2-3), 337-354.

Richards, W. (2008). Anigrafs: experiments in collective consciousness. Retrieved from http://people.csail.mit.edu/whit/contents.html.

Richards, W. (2008b). Modal Inference. In *Association for the Advancement of Artificial Intelligence Symposium*.

Richards, W., McKay, B. D., & Richards, D. (2002). The Probability of Collective Choice with Shared Knowledge Structures. *Journal of Mathematical Psychology*, *46*(3), 338-351. doi: 10.1006/jmps.2001.1391.

Richards, W., (2009). 9.343 Cognitive Architectures, course notes

Roberts, C. W. (1997). *Text analysis for the social sciences*. Routledge.

Romney, A. K., Weller, S. C., & Batchelder, W. H. (1986). Culture as Consensus: A Theory of Culture and Informant Accuracy. *American Anthropologist*, *88*(2), 313-338.

Rosen-Zvi, M., Griffiths, T., Smyth, P., & Steyvers, M. (2005, November 4). Learning Author Topic Models from Text Corpora. Retrieved May 14, 2009, from http://www.datalab.uci.edu/papers/UCI_KD-D_author_topic_preprint.pdf.

Rosen-Zvi, M., Griffiths, T., Steyvers, M., & Smyth, P. (2004). The author-topic model for authors and documents. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence* (pp. 487-494). Banff, Canada: AUAI Press. Retrieved January 29, 2009, from http://portal.acm.org/citation.cfm?id=1036843.1036902.

Rubinstein, A. (1982). Perfect Equilibrium in a Bargaining Model. *Econometrica*, *50*(1), 97-109.

Sackett, David L, William M C Rosenberg, J A Muir Gray, R Brian Haynes, and W Scott Richardson. 1996. Evidence based medicine: what it is and what it isn't. *BMJ* 312, no. 7023 (January 13): 71-72.

Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A Simplest Systematics for the Organization of Turn-Taking for Conversation. *Language*, *50*(4), 696-735.

Sah, R. K., & Stiglitz, J. E. (1988). Committees, Hierarchies and Polyarchies. *The Economic Journal*, *98*(391), 451-470.

Sapir, E., & Mandelbaum, D. G. (1949). *Selected writings of Edward Sapir*. University of California Press.

Shapiro, M., and D. Alighieri. 1990. *De vulgari eloquentia*. U of Nebraska Press, January 1.

Sherman, L. A. (2004). Looking Through a Window of the Food and Drug Administration: FDA's Advisory Committee System. *Preclinica*, *2*(2), 99-102.

Simon, H. A. (1964). On the Concept of Organizational Goal. *Administrative Science Quarterly*, *9*(1), 1-22.

Skvoretz, J. (1981). Extending Expectation States Theory: Comparative Status Models of Participation in N Person Groups. *Social Forces*, *59*(3), 752-770. doi: 10.2307/2578192.

Skvoretz, J. (1988). Models of Participation in Status-Differentiated Groups. *Social Psychology Quarterly*, *51*(1), 43-57. doi: 10.2307/2786983.

Smith-Lovin, L., Skvoretz, J. V., & Hudson, C. G. (1986). Status and Participation in Six-Person Groups: A Test of Skvoretz's Comparative Status Model. *Social Forces*, *64*(4), 992-1005. doi: 10.2307/2578790.

Stasser, G. (1988). Computer simulation as a research tool: The DISCUSS model of group decision making. *Journal of Experimental Social Psychology*, *24*(5), 393-422.

Stasser, G. (1992). Information salience and the discovery of hidden profiles by decision-making groups: A" thought experiment. *Organizational Behavior and Human Decision Processes*, *52*(1), 156-181.

Stasser, G., & Titus, W. (1985). Pooling of unshared information in group decision making: Biased information sampling during discussion. *Journal of Personality and Social Psychology*, *48*(6), 1467-1478.

Stasser, G., & Titus, W. (1987). Effects of information load and percentage of shared information on the dissemination of unshared information during group discussion. *Journal of Personality and Social Psychology*, *53*(1), 81-93.

Stasser, G., Stewart, D. D., & Wittenbaum, G. M. (1995). Expert roles and information exchange during discussion: The importance of knowing who knows what. *Journal of experimental social psychology*, *31*(3), 244-265.

Stephan, F. F., & Mishler, E. G. (1952). The Distribution of Participation in Small Groups: An Exponential Approximation. *American Sociological Review*, *17*(5), 598-608.

Tetlock, P. E. (2003). Thinking the unthinkable: sacred values and taboo cognitions. *Trends in Cognitive Sciences*, *7*(7), 320-324. doi: 10.1016/S1364-6613(03)00135-9.

Thomas-Hunt, M. C., Ogden, T. Y., & Neale, M. A. (2003). Who's really sharing? Effects of social and expert status on knowledge exchange within groups. *Management science*, *49*(4), 464-477.

Trice, H. M. (1993). *Occupational subcultures in the workplace*. Cornell University Press.

Visser, B., & Swank, O. H. (2007). On Committees of Experts. *Quarterly Journal of Economics*, 337-372.

Wallach, H. M. (2008). *Structured Topic Models for Language*. Doctor of Philosophy, University of Cambridge. Retrieved from http://www.cs.umass.edu/~wallach/theses/wallach_phd_thesis.pdf.

Wallach, H. M., & McCallum, D. M. (2009). Rethinking LDA: Why Priors Matter. *Topic Models: Text and Beyond Workshop in Neural Information Processing Systems Conference*, Whistler, BC.

Wang, X., Mohanty, N., & McCallum, A. (2005). Group and Topic Discovery from Relations and Text (pp. 28-35). Chicago, IL, USA: ACM.

Wasserman, S., & Faust, K. (1994). *Social Network Analysis: Methods and Applications* (1st ed.). Cambridge University Press.

Watanabe, S. (1985). *Pattern recognition: human and mechanical*. John Wiley & Sons, Inc. New York, NY, USA.

Wegner, D. M. (1987). Transactive memory: A contemporary analysis of the group mind. *Theories of group behavior*, 185-208.

Whorf, B. L., Carroll, J. B., & Chase, S. (1956). *Language, thought, and reality: Selected writings*. MIT press Cambridge, MA.

Winner, L. (1986). On not hitting the tar-baby. *The Whale and the Reactor: A Search for Limits in an Age of High Technology. Chicago*, *111*, 138-154.

Winquist, J. R., & Larson, J. R. (1998). Information Pooling: When It Impacts Group Decision Making. *Journal of Personality and Social Psychology*, *74*, 371-377.

Appendix 1

LIST OF FUNCTION WORDS (STOP LIST)

's	almost	astride
i	along	at
a	alongside	atop
aboard	altho	avec
about	although	away
above	amid	back
across	amidst	be
after	among	because
afterwards	amongst	before
against	an	beforehand
agin	and	behind
ago	another	behynde
agreed-upon	any	below
ah	anyone	beneath
alas	anything	beside
albeit	around	besides
all	as	between
all-over	aside	bewteen

beyond
bi
both
but
by
ca.
de
des
despite
do
down
due
durin
during
each
eh
either
en
every
ever
everyone
everything
except
far
fer
for
from
go
goddamn
goody
gosh
half
have
he
hell
her
herself
hey
him
himself
his
ho
how
however
i
if
in
inside
insofar
instead
into
it
its
itself
la
le
les
lest
lieu
like
me
minus
moreover

my
myself
near
near-by
nearer
nearest
neither
nevertheless
next
no
nor
not
nothing
notwithstanding
o
o'er
of
off
on
once
one
oneself
only
onto
or
other
others
otherwise
our
ours
ourselves
out
outside
outta
over
per
rather
regardless
round
se
she
should
since
so
some
someone
something
than
that
the
their
them
themselves
then
there
therefore
these
they
thine
this
those
thou
though

through
throughout
thru
till
to
together
toward
towardes
towards
uh
under
underneath
unless
unlike
until
unto
up
upon
uppon
us
via
vis-a-vis
vis-à-vis
we
well
what
whatever
whatsoever
when
whenever
where
whereas
wherefore
whereupon
whether
which
whichever
while
who
whoever
whom
whose
why
with
withal
within
without
ye
yea
yeah
yes
yet
yonder
you
your
yours
yourself
yourselves

Appendix 2

IMPACT OF DEMOGRAPHIC VARIABLES ON H-INDEX

We identify the effect of demographic variables on h-Index, a measure of academic expertise independent of panel dynamics. Results of this analysis are found in Table 30.

Table 30: 4-way ANOVA showing the effects of Age, Medical Specialty, Gender, and Race on h-index for our sample of 37 meetings. All variables reach significance.

Variable	Sum of Squares	Degrees of Freedom	Mean Squares	F	p-value
Age	8528.97	1	8528.97	51.79	<0.0001
Medical Specialty	23317.9	7	3331.12	20.23	<0.0001
Gender	2381.26	1	2381.26	14.46	0.0002
Race	2356.5	3	785.49	4.77	0.0029
Error	55333.8	336	164.68		
Total	98491.4	348			

These results show that age, medical specialty, gender, and race all contribute significantly to explaining the variance in h-index. Independent Tukey honestly-significant difference tests of multiple comparisons show that men have a significantly higher h-index than do women, and that white panel members have a significantly higher h-index than do black panel members. These results are known trends in the h-index, and are consistent with the analysis in (Kelly and Jennions 2006).

Appendix 3

CATALOGUE OF MEETINGS STUDIED

Meeting ID	Meeting Date	Device Name
1	'7/28/1997'	PLC CO2 Heart Laser
2	'7/29/1997'	Spectranetics Laser Sheath
3	'9/15/1997 - morning'	Alliance Monostrut Valve
4	'9/15/1997 - afternoon '	Medtronic Freestyle Aortic Root Bioprosthesis
5	'9/16/1997'	Toronto SPV® Valve, Model SPA-101
6	'4/24/1998'	PLC CO2 Heart Laser
7	'6/29/1998'	Ambu CardioPump
8	'10/27/1998'	Eclipse Holmium Laser
9	'6/23/1999 - morning'	Guidant Endovascular Technologies, EBT' Abdominal Aortic Tube Bifurcated Endovascular Grafting System
10	'6/23/1999 - afternoon'	Medtronic AneuRx, Inc. AneuRx Bifurcated Endovascular Prosthesis System

11	'6/24/1999'	Medtronic Jewel AF Arrhythmia Management Device
12	'6/19/2000'	Cordis Checkmate System
13	'9/11/2000'	Beta-Cath™ System intravascular brachytherapy device
14	'12/5/2000'	Model 7250 Jewel® AF Implantable Cardioverter Defibrillator System
15	'4/23/2001'	Sulzer IntraTherapeutics IntraCoil Self-Expanding Peripheral Stent
16	'7/9/2001'	Eclipse PMR Holmium Laser System
17	'7/10/2001 - morning'	Guidant Corporation P010012, Contak CD, and EasyTrak Lead System
18	'7/10/2001 - afternoon'	Medtronic Corporation P010015, Medtronic InSync Atrial Synchronous Biventricular Pacing Device and Attain Lead System
19	'9/10/2001 - morning'	AMPLATZER® Septal Occluder
20	'9/10/2001 - afternoon'	The CardioSEAL® Septal Occlusion System with QwikLoad™
21	'9/11/2001'	CryoLife, Inc. P010003, BioGlue Surgical Adhesive
22	'3/4/2002'	Thoratec Corporation's HeartMate VE Left Ventricular System
23	'3/5/2002'	InSync® ICD System
24	'9/9/2002'	Gore EXCLUDER® AAA Endoprosthesis

25	'9/10/2002'	NMT Medical P000049/S3, CardioSEAL STARFlex Septal Occlusion System with Qwik Load
26	'10/22/2002'	Cordis Corporation P020026, CYPHER Sirolimus-Eluting Coronary Stent System
27	'3/6/2003'	CryoCath Technologies' 7 French Freezor Cardiac Cryoablation Catheter
28	'4/10/2003'	Cook Zenith AAA Endovascular Graft
29	'5/29/2003'	Cardima, Inc. REVELATION_ Tx and NavAblator Catheter System
30	'10/2/2003'	Spectranetics CVX-300 Excimer Laser System
31	'11/20/2003'	TAXUS ® Drug Eluting Stent
32	'4/21/2004'	Cordis Precise Nitinol Stent System
33	'6/8/2004'	World Heart Novacor N100PC and N100PC(q) left ventricular assist system
34	'7/28/2004'	Guidant Cardiac Resynchronization Therapy Defibrillators, P010012, Supplement 26
35	'1/13/2005'	GORE TAG Thoracic Endoprosthesis
36	'6/22/2005'	Acorn CorCap™ Cardiac Support Device (CSD)
37	'6/23/2005'	Abiomed, Inc. H040006: AbioCor Implantable Replacement Heart

Meeting ID	Specialty Cohesion	Specialty Cohesion Percentile	Vote Cohesion	Vote Cohesion Percentile	Minority Size	Voting Outcome	Proportion Approving the Device
1	0.4	0.859	0.5333	0.541	0.22	0.00	0.22
2	0.2	0.566	N/A	N/A	0.00	1.00	1.00
3	0.3684	0.829	N/A	N/A	0.00	1.00	1.00
4	0.4545	0.883	N/A	N/A	0.00	1.00	1.00
5	0.625	0.966	N/A	N/A	0.00	1.00	1.00
6	0.5	0.832	N/A	N/A	0.00	1.00	1.00
7	0.2222	0.922	0.5	0.899	0.33	0.00	0.33
8	0.3636	0.975	N/A	N/A	0.14	1.00	0.86
9	0.25	0.507	N/A	N/A	0.00	1.00	1.00
10	0.3846	0.882	N/A	N/A	0.00	1.00	1.00
11	1	0.644	N/A	N/A	0.00	1.00	1.00
12	0.1765	0.913	N/A	N/A	0.00	1.00	1.00
13	0	0.713	N/A	N/A	0.00	1.00	1.00
14	1	0.968	N/A	N/A	0.00	1.00	1.00

15	0.2727	0.797	N/A	N/A	0.00	0.00	0.00
16	0.6667	0.959	1	1	0.22	0.00	0.22
17	0.25	0.534	0.625	0.667	0.25	0.00	0.25
18	0.1429	0.326	N/A	N/A	0.00	1.00	1.00
19	0.2667	0.792	N/A	N/A	0.00	1.00	1.00
20	0.25	0.735	N/A	N/A	0.10	1.00	0.90
21	1	0.983	N/A	N/A	0.00	1.00	1.00
22	0.3125	0.614	0.6875	0.674	0.20	1.00	0.80
23	0.2857	0.537	0.3571	0.487	0.50	1.00	0.50
24	0.2308	0.44	N/A	N/A	0.10	1.00	0.90
25	0.08	0.058	N/A	N/A	0.14	0.00	0.14
26	0.2	0.362	N/A	N/A	0.00	1.00	1.00
27	0.6364	0.996	0.8182	0.987	0.27	1.00	0.73
28	0.3333	0.53	N/A	N/A	0.00	1.00	1.00
29	1	0.967	N/A	N/A	0.00	0.00	0.00
30	0.0769	0.194	N/A	N/A	0.10	0.00	0.10

31	0.3846	0.862	N/A	N/A	0.00	1.00	1.00
32	0.2	0.774	0.6	0.911	0.40	1.00	0.60
33	0.2143	0.436	N/A	N/A	0.09	0.00	0.09
34	0.0909	0.302	N/A	N/A	0.00	1.00	1.00
35	0.7	0.989	0.7	0.949	0.20	1.00	0.80
36	0.2632	0.494	0.4737	0.277	0.36	0.00	0.36
37	0.1667	0.087	0.3667	0.586	0.54	1.00	0.46

Meeting ID	Length (Hours)	Author Topics	LDA Topics with Hyperparameter Optimization	Maximum Minority Outdegree	Committee Chair
1	7.67	28	232	0.3	1
2	2.67	9	148	N/A	1
3	3.00	13	152	N/A	2
4	5.10	14	129	N/A	2
5	3.10	16	158	N/A	2

6	7.25	19	341	N/A	2
7	9.17	23	259	0.3333	2
8	8.00	18	207	0	3
9	3.92	14	217	N/A	2
10	4.45	9	155	N/A	2
11	3.67	14	165	N/A	3
12	7.08	21	304	N/A	2
13	7.08	20	292	N/A	4
14	5.98	24	203	N/A	4
15	8.08	17	212	N/A	4
16	7.35	20	300	0	4
17	4.65	16	173	0	1
18	3.37	11	135	N/A	1
19	3.93	12	128	N/A	4
20	5.00	14	128	0.2222	4
21	2.97	11	155	N/A	5

22	7.67	18	260	0.2222	5
23	7.42	18	236	0.4	5
24	7.92	23	228	0	5
25	7.25	18	263	0.0833	4
26	11.62	22	302	N/A	5
27	7.42	19	244	0.1	5
28	7.03	28	207	N/A	5
29	7.55	20	269	N/A	5
30	6.45	19	198	0	5
31	7.53	22	305	N/A	5
32	10.33	28	300	0.4	5
33	7.15	22	176	0.3	5
34	8.38	23	208	N/A	5
35	8.28	25	299	0.0909	6
36	9.70	21	295	0.3333	6
37	8.87	25	257	0.3571	6

Meeting ID	Number of Cycles No Chair	Number of Cycles with Chair	Added Cycles	Normalized Cycles	Cycle Proportion No Chair	Cycle Proportion with Chair	Chair Effect
1	14	17	3	0.18	0.70	0.68	-0.02
2	0	0	0	0.00	0.00	0.00	0.00
3	0	0	0	0.00	0.00	0.00	0.00
4	7	7	0	0.00	0.50	0.39	-0.11
5	10	16	6	0.38	0.83	0.94	0.11
6	4	11	7	0.64	0.50	0.85	0.35
7	2	2	0	0.00	0.11	0.09	-0.02
8	0	0	0	0.00	0.00	0.00	0.00
9	0	11	11	1.00	0.00	0.61	0.61
10	2	4	2	0.50	0.15	0.21	0.06
11	0	0	0	0.00	0.00	0.00	0.00
12	9	15	6	0.40	0.50	0.60	0.10

13	5	10	5	0.50	0.63	0.77	0.14
14	0	2	2	1.00	0.00	0.29	0.29
15	3	9	6	0.67	0.27	0.53	0.26
16	2	2	0	0.00	0.20	0.15	-0.05
17	0	0	0	0.00	0.00	0.00	0.00
18	6	6	0	0.00	0.67	0.50	-0.17
19	0	0	0	0.00	0.00	0.00	0.00
20	15	20	5	0.25	0.94	0.91	-0.03
21	0	0	0	0.00	0.00	0.00	0.00
22	0	7	7	1.00	0.00	0.33	0.33
23	0	0	0	0.00	0.00	0.00	0.00
24	0	7	7	1.00	0.00	0.35	0.35
25	5	19	14	0.74	0.19	0.56	0.37
26	0	0	0	0.00	0.00	0.00	0.00
27	0	7	7	1.00	0.00	0.41	0.41
28	2	7	5	0.71	0.29	0.50	0.21

29	3	7	4	0.57	1.00	1.00	0.00
30	9	20	11	0.55	0.53	0.87	0.34
31	13	13	0	0.00	0.68	0.57	-0.12
32	0	8	8	1.00	0.00	0.32	0.32
33	8	8	0	0.00	0.47	0.42	-0.05
34	12	12	0	0.00	0.80	0.52	-0.28
35	12	21	9	0.43	0.80	0.91	0.11
36	4	13	9	0.69	0.19	0.52	0.33
37	12	35	23	0.66	0.35	0.78	0.42