Appendix A: Comparison of automated text analysis and hand coded ties

In order to better understand the nature of the data generated by our automated text analysis, we hand-coded a sample of 49 randomly selected meetings for comparison. Figure A1 orders these 49 meetings along the X-axis by the number of attendees observed via hand-coding. For each of these meetings, we then plot the number of attendees observed via hand-coding (hollow, darkly bordered points) and the number of attendees observed via automated coding (solid, lightly colored points).

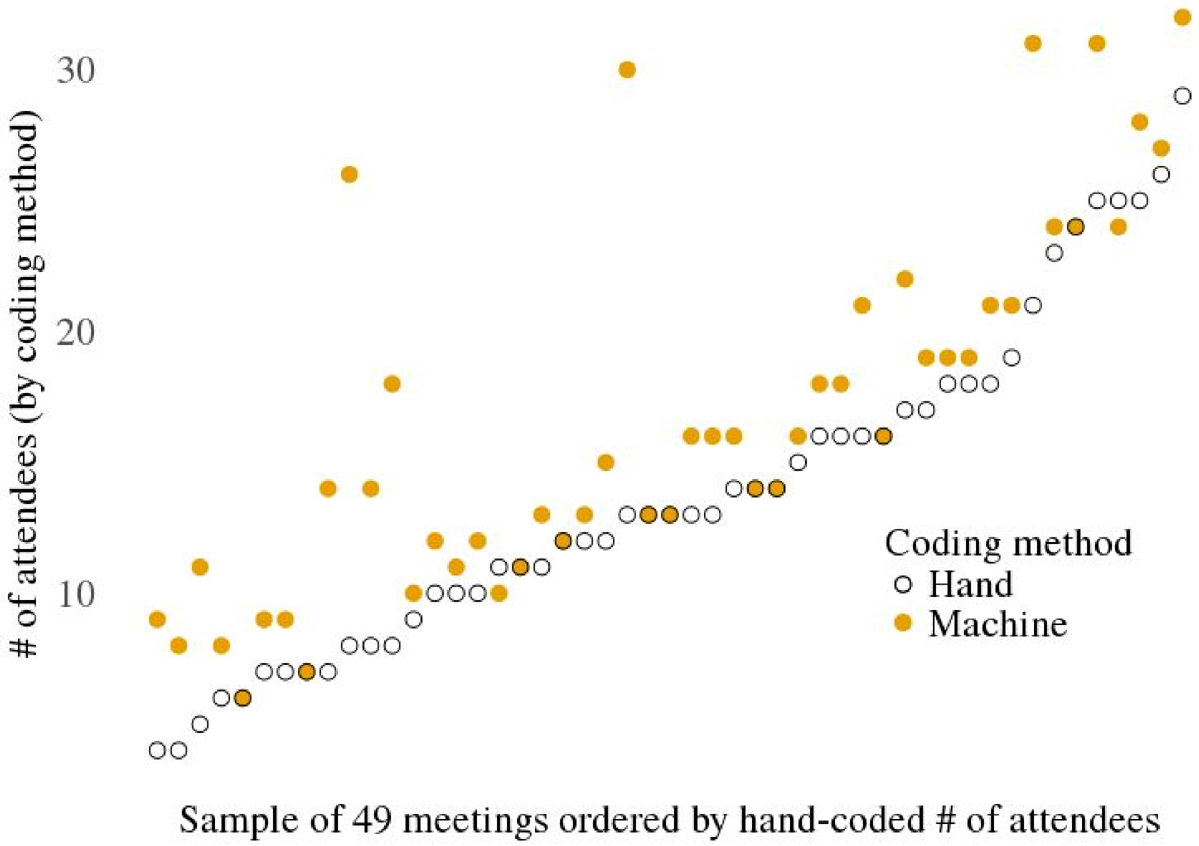

Figure A1: Comparison of hand-coded and machine-coded meeting attendance

Figure A1 shows that for the most part, automated attendance tracks fairly closely to hand-coded attendance. With two exceptions where the machine-coded number is one unit lower than the hand-coded number, the machine-coded data appear to be most subject to false positives. This is to be expected, since text processing algorithms are excellent at extracted named entities found in text (and thus in general not prone to overlook a name and generate a false negative), but contextualizing the reason why said name occurs in the text is more challenging. An examination of the cases shown in figure A1 where the machine-coded attendance number far outpaces the hand-coded number revealed that these meeting summary documents contained a both a list of those present at the meeting and a list of individuals on a “distribution list” who did not attend the meeting but desired to be kept apprised of the process. On a per-meeting basis, the average difference between hand-coded attendance and machine-coded attendance is 2.96 more machine coded attendees. As the average number of attendees per meeting within the 49 meeting subsample is 13.8, machine-coding on average over-inflates attendance by around 21%.

It is also important to note that the machine-coded attendance data are not purely the raw extracted named entities. After performing entity extraction (as described in the methods section above), we took a series of basic data cleaning steps (the code for which are available as part of the data and code for this analysis made publicly available). These steps included filtering out obvious false positives that did not correspond to actual names. For example, terms such as “Dolly Varden” (a trout species native to the region) and “Howard Hansen” (the name of another dam in the region) are incorrectly recognized as named persons. Further, the structure of the meeting summary documents poses a challenge in many cases, for instance, where participation is itemized such that an entry might say something akin to: “Tony Revise document and provide new draft next month.” Given the capitalization of “Revise”, this is prone to being extracted as an entity named “Tony Revise”. Thus, we also filter out extracted entities that contain actions such as “revised”, “draft”, “email”, “clean”, and “report” (after carefully checking to ensure that any words used as filters do not also happen to be the name of a participating stakeholder). We also corrected for misspelled names with code script that used regular expression searches to identify pairs of extracted named entities that differed by a limited number of character changes, additions, or removals.

While attendance data are used to identify tie recipients in the model, participation data are needed to code where ties originate. We conduct a similar comparison exercise using the hand-coded meeting sample. Figure A2 compares the number of hand-coded participation actions observed for each meeting with the number of participation actions observed via our automated text analysis. For the most part, we again see that machine-coded observations do track reasonably well with hand-coded observations. However, whereas automated coding of attendance primarily generated false positives, figure A2 shows that automated coding of participation appears more prone to producing false negatives (i.e., failing to recognize a stakeholder’s participation action). The most prominent driver of this is that that hand-coding is better able to disambiguate names and to identify the subject(s) of pronouns. In other words, a contextualized reading of a document increases the chance that a recognized participation action can be clearly attributed to a specific attendee. Thus, the machine-coding approach exhibits a downward bias in terms of observing participation.

On a per-meeting basis, the average difference between the number of hand-coded participation actions and machine-coded participation actions is 4.3 more hand-coded observations. The average number of total participation actions observed per meeting in the hand-counted sample is 22.73; machine-coding thus captures about 81% of total participation per meeting on average.

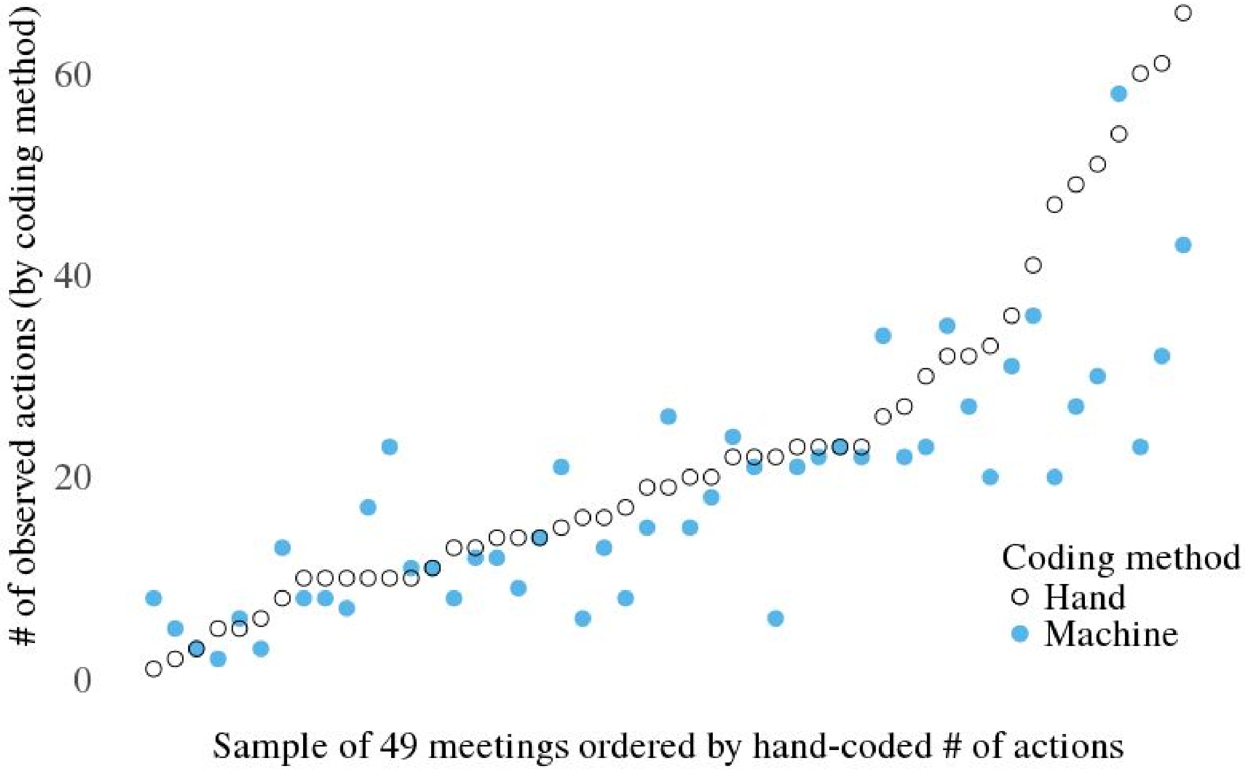

Figure A2: Comparison of hand-coded and machine-coded meeting participation actions

While figure A2 presents the total count of all participation actions observed for each meeting via hand-coding and automated-coding methods, recall that network ties are coding in binary fashion for this analysis. In other words, if an attendee is documented in a meeting summary to have given a presentation, provided a comment, and made a suggestion (i.e., three separate participation actions), they are only assigned one "interaction" tie stemming to other attendees. Thus, with respect to understanding how the coding method used might influence analysis results, it is important to consider how the hand-coded and automatically coded data differ with respect to the number of unique stakeholders observed to participate in a given meeting. In other words, for the current analysis is it does not really matter if the hand-coded and automatically coded participation data differ in terms of how many times an attendee participation; rather, what can influence the results is whether an attendee is recorded as having participated at all.

In terms of capturing unique individual meeting participants, the automated coding strategy performs very well. Figure A3 compares hand-coded and automatically coded unique participants by meeting; while it might appear that there is considerable divergence, note that the y-axis is shrunken considerably relative to figure A2. No individual meeting differs by more than 5 observations between the two codings, and the average difference between hand-coded and machine-coded observations is 0.38 observations per meeting. The average number of actions taken by unique stakeholders per meeting is 6.35, meaning that machine coding captures about 94% of unique participation on average.

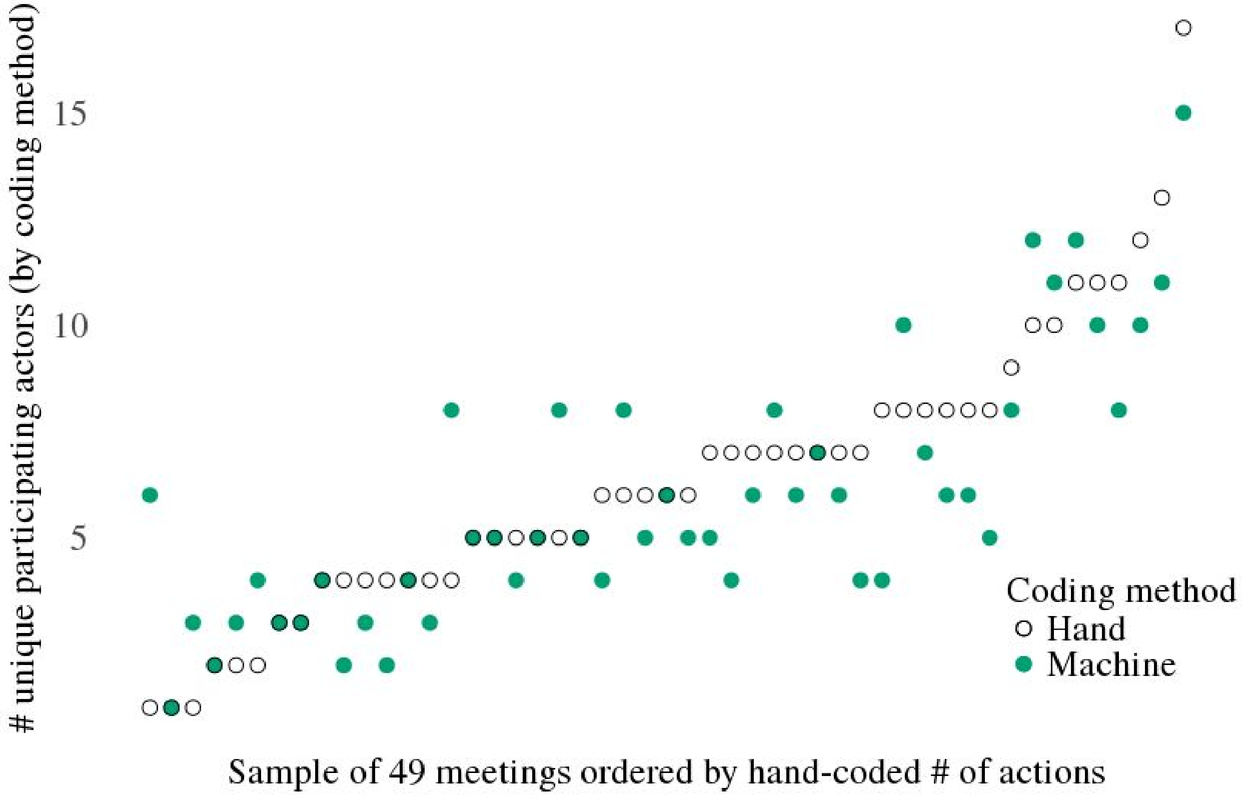

Figure A2: Comparison of hand-coded and machine-coded participating meeting actors