As Long as You Name My Name Right: Social Circles and Social Sentiment in the Hollywood Hearings

Oren Tsur†§ Dan Calacci† David Lazer†
orentsur@seas.harvard.edu dcalacci@ccs.neu.edu dlazer@neu.edu
†Lazer Laboratory, Northeastern University
§School of Engineering and Applied Sciences, Harvard University

Abstract
The Hollywood Blacklist was based on a series of interviews conducted by the House Committee on Un-American Activities (HUAC), trying to identify members of the communist party. We use various NLP algorithms in order to automatically analyze a large corpus of interview transcripts and construct a network of the industry members and their “naming” relations. We further use algorithms for Sentiment Analysis in order to add a psychological dimension to the edges in the network. In particular, we test how different types of connections are manifested by different sentiment types and attitude of the interviewees. Analysis of the language used in the hearings can shed new light on the motivation and role of network members.

1 Introduction
A growing body of computational research is focused on how language is used and how it shapes/is shaped by a community of speakers. Computational works in the nexus of language and the social arena deal with various topics such as language accommodation (Danescu-Niculescu-Mizil and Lee, 2011; Danescu-Niculescu-Mizil et al., 2011), demographic language variation (Eisenstein et al., 2010; O’Connor et al., 2010), the factors that facilitate the spread of information in Q&A forums and social networks (Adamic et al., 2008; Bian et al., 2009; Romero et al., 2011) or the correlation between words and social actions (Adali et al., 2012).

All of these works analyze the language and the social dynamics in online communities, mainly due to the increasing popularity of online social networks and greater availability of such data.

However, large scale socio-linguistic analysis should not be restricted to online communities and can be applied in many social and political settings beyond the online world. Two examples are the study of power structures in arguments before the U.S. Supreme Court (Danescu-Niculescu-Mizil et al., 2012) and the evolution of specific words and phrases over time as reflected in Google Books (Goldberg and Orwant, 2013).

In this paper we propose using network science and linguistic analysis in order to understand the social dynamics in the entertainment industry during one of its most controversial periods – the ‘red scare’ and the witch hunt for Communists in Hollywood during 1950’s.

Historical background The Hollywood hearings (often confused with Senator McCarthy’s hearings and allegations) were a series of interviews conducted by the House Committee on Un-American Activities (HUAC) in the years 1947–1956. The purpose of the committee was to conduct “hearings regarding the communist infiltration of the motion picture industry” (from the HUAC Annual Report). The committee subpoenaed witnesses such as Ayn Rand (writer), Arthur Miller (writer), Walt Disney (producer), future U.S. president Ronald Reagan (Screen Actors Guild), Elia Kazan (writer, actor, director) and Albert Maltz (Screen Writers Guild). Some of the witnesses were ‘friendly’ while some others were uncooperative, refusing to “name names” or self incriminate. Those who were named and/or were uncooperative were often jailed or effectively lost their job.

Arguably, many friendly witnesses felt they were complying with their patriotic duty. Many

---

1 A note about terminology: by using the terms friendly and uncooperative there is no implied moral judgment – these are the terms used in the literature.

2 It should be noted that being a member of the Communist party was not illegal, however, some individuals avoided self “incrimination” either in an effort to protect their job or as an ideological declaration in favor of privacy protection as a civil right protected by the constitution.
others were threatened or simply manipulated to name names, and some later admitted to cooperating for other reasons such as protecting their work or out of personal vendettas and professional jealousies. It is also suspected that some naming occurred due to increasing professional tension between some producers and the Screen Writers Guild or (Navasky, 2003).

**Motivation** In this work we analyze a collection of HUAC hearings. We wish to answer the following questions:

1. Do sentiment and other linguistic categories correlate with naming relations?
2. Can we gain any insight on the social dynamics between the people in the network?
3. Does linguistic and network analysis support any of the social theories about dynamics at Hollywood during that time?

In order to answer the questions above we build a social graph of members of the entertainment industry based on the hearings and add sentiment labels on the graph edges. Layering linguistic features on the social graph may provide us with new insights related to the questions at hand. In this short paper we describe the research framework, the various challenges posed by the data and present some initial promising results.

**2 Data**

In this work we used two types of datasets: Hearing Transcripts and Annual Reports. Snippets from hearings can be found in Figures 1(a) and 1(b). Figure 1(c) shows a snippet from an annual report. The transcripts data is based on 47 interviews conducted by the HUAC in the years 1951–2. Each interview is either a long statement (1(a)) or a sequence of questions by the committee members and answers by a witness (1(b)). In total, our hearings corpus consists of 2831 dialogue acts and half a million words.

**3 Named Entity Recognition and Anaphora Resolution**

The snippets in Figure 1 illustrates some of the challenges in processing HUAC data. The first challenge is introduced by the low quality of the available documents. Due to the low quality of the documents the OCR output is noisy, containing misidentified characters, wrong alignment of sentences and missing words. These problems introduce complications in tasks like named entity recognition and properly parsing sentences.

Beyond the low graphic quality of the documents, the hearings present the researcher with the typical array of NLP challenges. For example, the hearing excerpt in 1(b) contains four dialogue acts that need to be separated and processed. The committee member (Mr. Tavenner) mentions the name Stanley Lawrence, later referred to by the witness (Mr. Ashe) as Mr. Lawrence and he thus coreference resolution is required before the graph construction and the sentiment analysis phases.

As a preprocessing stage we performed named entity recognition (NER), disambiguation and unification. For the approximately 19 months of my membership, I was assigned to a "unit" composed of three part-time members who, like myself, members of the Group Theatre acting company. These were—

Levi Levitt, co-founder of the unit.
J Edward Reinking, co-founder of the unit, deceased.
Whitmore Brown, sister Mrs. Sherry Oates.
I was instrumental in bringing her into the party.
Sherri Oates
Troy Kreisinger, along with William (two below), he recruited me into the party.
Paul Miller (Mrs. Lee Sherrill)
We are friends today. I believe that she also has told me she got the Communist bug—she is far too sensible and balanced a woman, and she is married to too fine an intellectual a man, to have remained among them.

Clifford Odets. He has assured me that he got out about the same time I did. Art Smith.
We used the Stanford Deterministic Coreference Resolution System (Lee et al., 2011) to resolve anaphoric references.

4 Naming Graph vs. Mentions Graph

In building the network graph of the members of the entertainment industry we distinguish between mentioning and naming in our data. While many names may be mentioned in a testimony (either by a committee member or by the witness, see example in Figures 1(a) and 1(b)), not all names are practically ‘named’ (=identified) as Communists. We thus use the hearings dataset in order to build a social graph of mentions (MG) and the annual reports are used to build a naming graph (NG). The NG is used as a “gold standard” in the analysis of the sentiment labels in the MG. Graph statistics are presented in Table 1.

While the hearings are commonly perceived as an “orgy of informing” (Navasky, 2003), the difference in network structure of the graphs portrays a more complex picture. The striking difference in the average out degree suggests that while many names were mentioned in the testimonies (either in a direct question or in an answer) – majority of the witnesses avoided mass-explicit naming. The variance in outdegree suggests that most witnesses did not cooperate at all or gave only a name or two, while only a small number of witnesses gave a long list of names. These results are visually captured in the intersection graph (Figure 2) and were also manually verified.

The difference between the MG and the NG graph in the number of nodes with out-going edges (214 vs. 66) suggests that the HUAC used other informers that were not subpoenaed to testify in a hearing.

In the remainder of this paper we analyze the the distribution of the usage of various psychological categories based on the role the witnesses play.

5 Sentiment Analysis and Psychological Categories

5.1 Sentiment Analysis

We performed the sentiment analysis in two different settings: lexical and statistical. In the lexical setting we combine (Ding et al., 2008) and the LIWC lexicon (Tausczik and Pennebaker, 2010). In the statistical setting we use NaSent (Socher et al., 2013).

The motivation to use both methods is twofold: first – while statistical models are generally more robust, accurate and sensitive to context, they require parsing of the processed sentences. Parsing our data is often problematic due to the noise introduced by the OCR algorithm due to the poor quality of the documents (see Figure 1). We expected the lexicon-based method to be more tolerant to noisy or ill-structured sentences. We opted for the LIWC since it offers an array of sentiment and psychological categories that might be relevant in the analysis of such data.

Table 1: Network features of the Mentions graph, the Naming graph and the intersection of the graphs.

<table>
<thead>
<tr>
<th></th>
<th>MG</th>
<th>NG</th>
<th>Intersection</th>
</tr>
</thead>
<tbody>
<tr>
<td>Num of nodes</td>
<td>1353</td>
<td>631</td>
<td>122</td>
</tr>
<tr>
<td>Num of edges</td>
<td>2434</td>
<td>842</td>
<td>113</td>
</tr>
<tr>
<td>Nodes / Edges</td>
<td>0.55</td>
<td>0.467</td>
<td>1</td>
</tr>
<tr>
<td>Avg. out degree</td>
<td>36.87</td>
<td>9.93</td>
<td>8.7</td>
</tr>
<tr>
<td>Avg. in degree</td>
<td>1.82</td>
<td>1.83</td>
<td>1.04</td>
</tr>
<tr>
<td>Var(outdegree)</td>
<td>3902.62</td>
<td>120.75</td>
<td>415.59</td>
</tr>
<tr>
<td>Var(indegree)</td>
<td>4.0</td>
<td>2.51</td>
<td>1.04</td>
</tr>
<tr>
<td>Nodes with out going edges</td>
<td>66</td>
<td>214</td>
<td>13</td>
</tr>
<tr>
<td>Nodes with incoming edges</td>
<td>1341</td>
<td>459</td>
<td>109</td>
</tr>
<tr>
<td>Reciprocity</td>
<td>0.016</td>
<td>0.012</td>
<td>0</td>
</tr>
</tbody>
</table>

Figure 2: Naming graph based on the intersection of the mentions and the naming data. Larger node size indicates a bigger out degree; Color indicates the in degree (darker nodes were named more times).

3 Ayn Rand and Ronald Reagan, two of the most ‘friendly’ witnesses (appeared in front of the HUAC in 1947), did not name anyone.

4 There might be some hearings and testimonies that are classified or still not publicly accessible.
Aggregated Sentiment A name may be mentioned a number of times in a single hearing, each time with a different sentiment type or polarity. The aggregated sentiment weight of a witness $i$ toward a mentioned name $j$ is computed as follows:

$$sentiment(i, j) = \max_{c \in CAT} \frac{\sum_{k \in U_{ij}} \text{score}(u_{ij}^k, c)}{|U_{ij}|}$$

Where $CAT$ is the set of categories used by LIWC or Stanford Sentiment and $U_{ij}$ is the set of all utterances (dialogue acts) in which witness $i$ mentions the name $j$. The $\text{score}(\cdot)$ function is defined slightly different for each setting. In the LIWC setting we define $\text{score}(\cdot)$ as:

$$\text{score}(u_{ij}^k, c) = \frac{|\{w \in u_{ij}^k | w \in c\}|}{|u_{ij}^k|}$$

In the statistical setting, Stanford Sentiment returns a sentiment category and a weight, we therefore use:

$$\text{score}(u_{ij}^k, c) = \begin{cases} w_c, & \text{if sentiment found} \\ 0, & \text{if } c \text{ was not returned} \end{cases}$$

Unfortunately, both approaches to sentiment analysis were not as useful as expected. Most graph edges did not have any sentiment label, either due to the limited sentiment lexicon of the LIWC or due to the noise induced in the OCR process, preventing the Stanford Sentiment engine from parsing many of the sentences. Interestingly, the two approaches did not agree on most sentences (or dialogue acts). The sentiment confusion matrix is presented in Table 2, illustrating the challenge posed by the data.

5.2 Psychological Categories

The LIWC lexicon contains more than just positive/negative categories. Table 3 presents a sample of LIWC categories and associated tokens. Figure 3 presents the frequencies in which each psychological category is used by friendly and uncooperative witnesses. While the Pronoun category is equally used by both parties, the uncooperative witnesses tend to use the I, Self and You categories while the friendly witnesses tend to use the Other and Social. A somewhat surprising result is that the Tentat category is used more by friendly witnesses – presumably reflecting their discomfort with their position as informers.

6 Conclusion and Future Work

In this short paper we take a computational approach in analyzing a collection of HUAC hearings. We combine Natural Language Processing and Network Science techniques in order to gain a better understanding of the social dynamics within the entertainment industry in its darkest time. While sentiment analysis did not prove as useful as expected, analysis of network structures and the language usage in an array of psychological dimensions reveals differences between friendly and uncooperative witnesses.

Future work should include a better preprocessing of the data, which is also expected to improve the sentiment analysis. In future work we will analyze the language use in a finer granularity of witness categories, such as the ideological informer, the naive informer and the vindictive informer. We also hope to expand the hearings corpora to include testimonies from more years.

Acknowledgements

We thank Giacomo Negro and Sasha Goodman for generously sharing their HUAC hearings transcripts and reports.
References


Yoav Goldberg and Jon Orwant. 2013. Syntactic ngrams over time from a very large corpus of english books. In Second Joint Conference on Lexical and Computational Semantics.


