Mining Text Networks:
A Cross-Disciplinary Science

David Banks
Duke University

1. Introduction

Text networks arise in many situations:

- the Wikipedia is a network whose nodes are articles and whose edges are hyperlinks; each article contains text;
- citation networks;
- Internet webpages.

One would like to use information in the text to improve network models of connectivity, or growth and change, and use network information to improve topic discovery.

Progress would enable one to find “holes” in the Wikipedia, improve recommender systems, and increase click-through rates for Internet advertisements.

A fair amount of previous work has been done in this area. The two main approaches involve:

- Natural language models
- Bag-of-words models.

The latter approach ignores semantic information: “I am not a crook” and “Am I not a crook” provide equivalent signal.

In contrast, natural language models attempt to include semantic information. In the context of text networks, there is a connection to the Semantic Web (cf. Allan Collins and Tim Berners-Lee), which attempts to provide hypertext metadata to provide “a web of data that can be processed directly and indirectly by machines” (Berners-Lee).

Natural language models are really hard. Oddly, the bag-of-words models are insanely successful for many purposes.

In practice, many researchers use n-grams to incorporate some semantic information into a bag-of-words analysis. An **n-gram** is a sequence of n word stems that has relatively high probability of co-occurring (e.g., “President of the United States”).

A **word stem** is a base word. The words “swam,” “swum,” “swim,” “swimming” all map to the same base. It is usually helpful to ignore tense, plurals, and other minor variations.

For example, if the first word is “how” then with high probabability the next word will be “are” or “can” or “is” or “will” or “do” or a handful of others. And the third word is, with fairly high probability, one of “you” or “we” or “I” or “one” or “my” or “your” or “Mom” and so forth. This ripple of excess probability flattens out to something close to baseline after about 8 or 9 words.

English has a window of about 8 to 9 before the Shannon entropy measure gets really high. Thus, after being told a specific word in a communication, the conditional probability of the eighth or ninth word after that is essentially the raw frequency of that word.

Latent Semantic Indexing is a procedure that addresses semantic problems of synonomy and polysemy by interpreting the meaning of words in the context of other words in the same document.

Synomyms are an issue for n-grams. Probability for the same “meaning” gets allocated across multiple sequences. But LSI can recognize synonyms:

- reduce the deficit by taxing job creators
- reduce the national debt by taxing fat cats

lead to the n-grams“job creators” and “fat cats” being nearby in term space.

Polysemy is harder; it requires disambiguations, and one wants to use only cues in the text, not domain knowledge, to do this.

For example, “Grateful Dead” can refer to a rock band or to a genre of German folktale. If the document includes the words “music” or “drugs” or “Haight-Ashbury” then the context suggests the former meaning. But if the document contains “woodcutter” or “coffin” or “magic goose” then the latter sense is implied.

LSI does singular value decomposition on a contingency table of text. This is sometimes called **correspondence analysis** (cf. Benzécri, 1973).

LSI starts with a term-document matrix $\boldsymbol{X}$. The rows consist of all n-grams in the corpus, the columns list all documents, and the cells contain counts. Then one normalizes for the relative frequency of n-gram within the document and the relative frequency of the word in the corpus.

The SVD finds matrices $\boldsymbol{T}$, $\boldsymbol{S}$ and $\boldsymbol{D}$ such that $\boldsymbol{X} = \boldsymbol{TSD}'$ where

- $\boldsymbol{S}$ is a diagonal matrix containing the singular values,
- $\boldsymbol{T}$ is the term matrix whose rows are eigenvectors that define the "term space",
- $\boldsymbol{D}$ is the document matrix whose columns are eigenvectors that define the "document space".

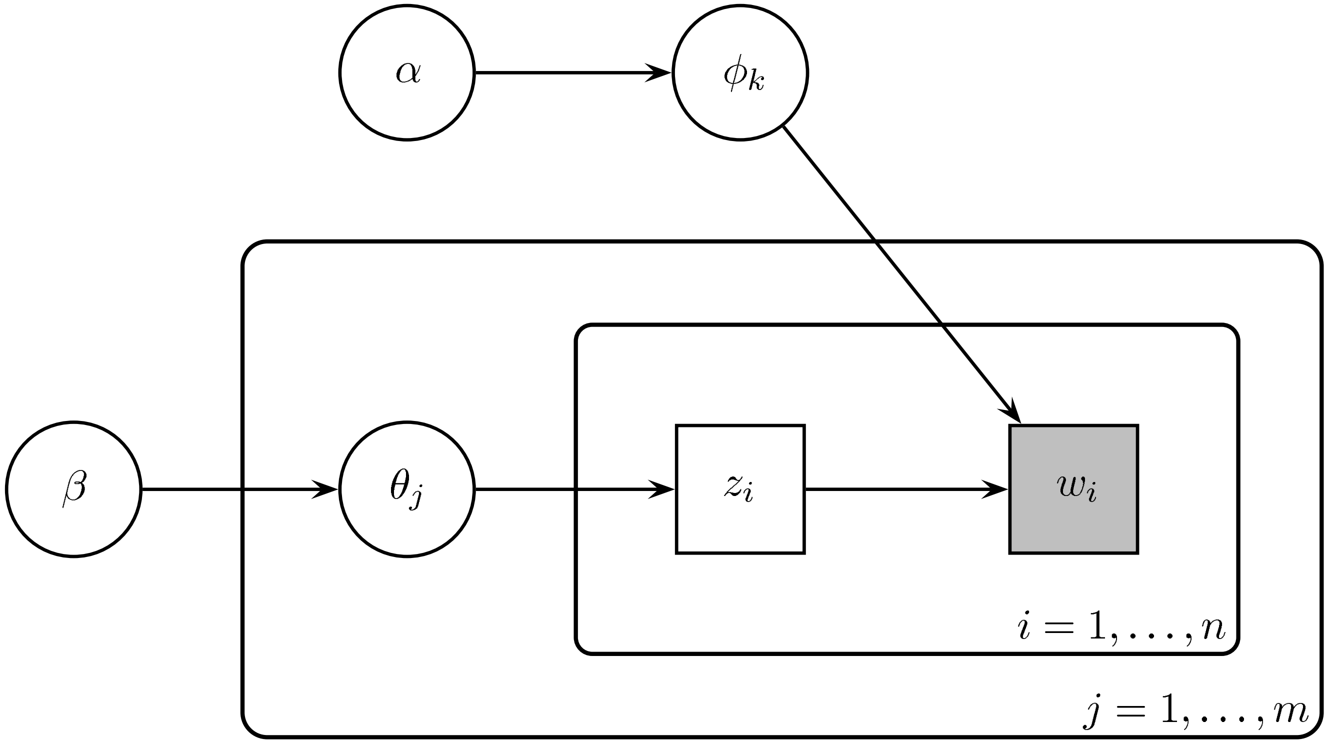
If n-grams are near in term space, they are synonyms; if they have clusters, they are polysemic. Usually it is good to truncate the $\boldsymbol{S}$ matrix.

Currently, topic models are the popular approach. Most use some form Latent Dirichlet Allocation (LDA), by Blei, Ng, and Jordan (2002). The generative model for assigning words to documents is as follows:

- For each topic, draw $\phi_k \in \mathbb{R}^v$ from a Dirichlet distribution with parameter $\boldsymbol{\alpha}$; this determines the distribution over the vocabulary for topic k.
- For document D_j, draw $\boldsymbol{\theta}_j \in \mathbb{R}^K$ from a Dirichlet distribution with parameter $\boldsymbol{\beta}$, which determines the extent to which document D_j participates in each of the K topics.
- Independently, for each word in document D_j, first draw a single topic z_i from the one-trial multinomial with parameter $\boldsymbol{\theta}_j$. If one draws the ith topic, then the word is chosen as a single draw from the one-trial multinomial with parameter $\boldsymbol{\phi}_i$.

LDA chooses topics for each document according to one Dirichlet distribution, and then, conditional on a topic, the vocabulary is chosen according to its corresponding Dirichlet distribution.

This generative model is often described through a plate diagram, which represents the relationships between the mechanisms for composing documents as random mixtures of topics, with vocabulary drawn independently, with probabilities depending upon the topic.



2. Political Blogs: A Case Study

The political blogosphere is a dynamic network of documents. Discussion on one site will respond to or reference comments in other documents, and there is a clear sense that important memes get passed around within the network.

To study this, we scraped the text from the 1,509 top U.S. political blogs (as determined by `Technorati`) between Jan. 1 2012 and Dec. 31 2012.

The importance of a political blog is estimated by `Technorati` based on the the site's "standing and relevance in the blogosphere" as determined by the relevance of its content and link behavior.

As a starting point, our analysis focused on blog posts that discussed the Trayvon Martin shooting on Feb. 26, 2012.

When dealing with diverse sources, one requires complementary skill sets. The research team included:

- Computer scientists at MaxPoint Interactive, a start-up in the Research Triangle that does computational advertising. They scraped and tokenized the text (after declaring robot status and following all robot protocols).
- Justin Gross and various political science students at UNC-Chapel Hill. They verified the `Technorati` ratings, recommended specific news events for examination, and hand-validated a random sample of the scraped text.
- Tim Au and David Banks, statisticians at Duke University. We fit topic models and network models.

The initial tokenization was problematic. The MaxPoint software removed all three-letter words (e.g., NRA, EPA, DHS, and so forth). And there was a good bit of back-and-forth initially, as various quality issues were sorted out.

As context, recall that Trayvon Martin was an unarmed 17-year-old African-American teenager who was shot by George Zimmerman in a gated community in Sanford, FL.

- Zimmerman was the neighborhood watch coordinator for the community. He thought Martin was acting suspiciously, and reported him to the police.
- The police instructed Zimmerman not to approach the suspect, but were disregarded.
- Zimmerman claims that Martin attacked him, and he used his gun in self-defense; initially, he was not charged with any crime.
- About 10 days after the shooting, the story exploded in the national media. The discussion included issues of racism, inaccurate and biased coverage, and legal questions.
- On May 3, Zimmerman was charged with second degree murder and pled not guilty.

Of the 114,611 blog entries that were scraped, 1,103 from 145 domains mention "Trayvon" one or more times. This was the corpus for analysis, together with links to made by those posts to any other of the 145 domains.

There were some data quality problems. Known issues include:

- Inconsistent dating protocols between blog sites.
- The scraped text did not match the on-line text ($\approx 8\%$ of the domains).
- Scraping captured reader comments ($\approx 27\%$ of the domains).

To validate, political science students rated a sample of posts for relevance and accuracy.

To begin the analysis of the text, we first identified the n-grams. The most common (tokenized) n-grams were:

georg zimmerman
year old
presid obama
look like
barack obama
fox news
trayvon martin
dont know
stand ground
unit state
dont think
hate crime
self defens
stand ground law
african american
trayvon martin case
law enforc
civil right

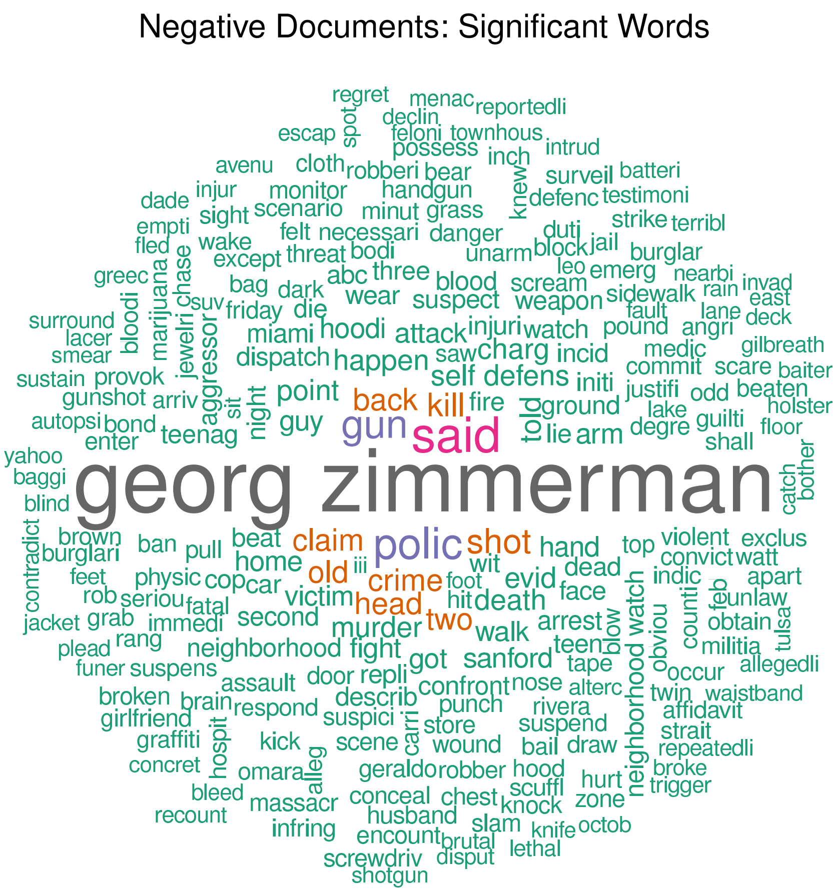
As a small digression before the network study, consider a sentiment analysis of the data. Sentiment analysis involves trying to determine whether a document has a positive, negative, or neutral tone. Irony and sarcasm make this hard—humans agree only about 80% of the time.

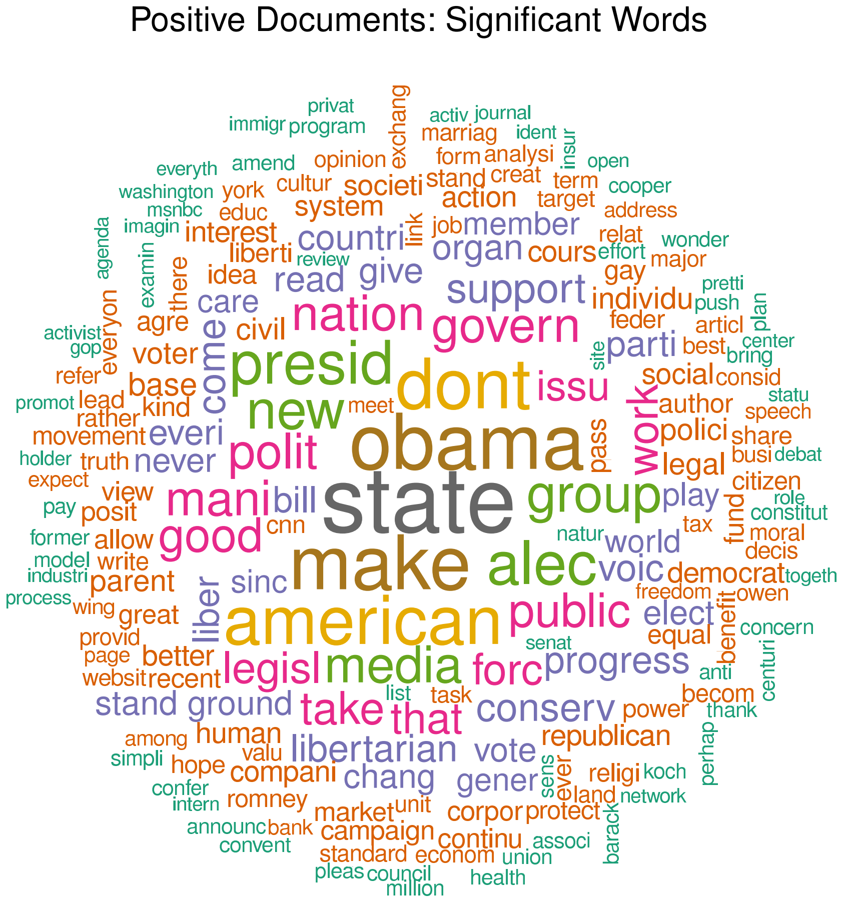
AFINN is a list of 2,477 English words scored from -5 to 5 in terms of positive/negative connotation (and negation reverses the sign). To reduce situational bias, before measuring the tones of the blog posts we removed terms that had negative association but which were necessarily relevant to the event; e.g., kill, gun, crime.

To measure the polarity of a document, let d_w be the number of occurrences of word w and let s_w be the AFINN score of that word. Then the polarity of a document is

$$\mathbf{polarity} = \frac{\sum d_w s_w}{\sum d_w |s_w|} \in [-1, 1]$$

We now identify the words in the blog posts that were differentially used between posts that had positive tone (i.e., polarity greater than 0.2) and those that had negative tone (polarity less than -0.2).


Negative Documents: Significant Words


Positive Documents: Significant Words

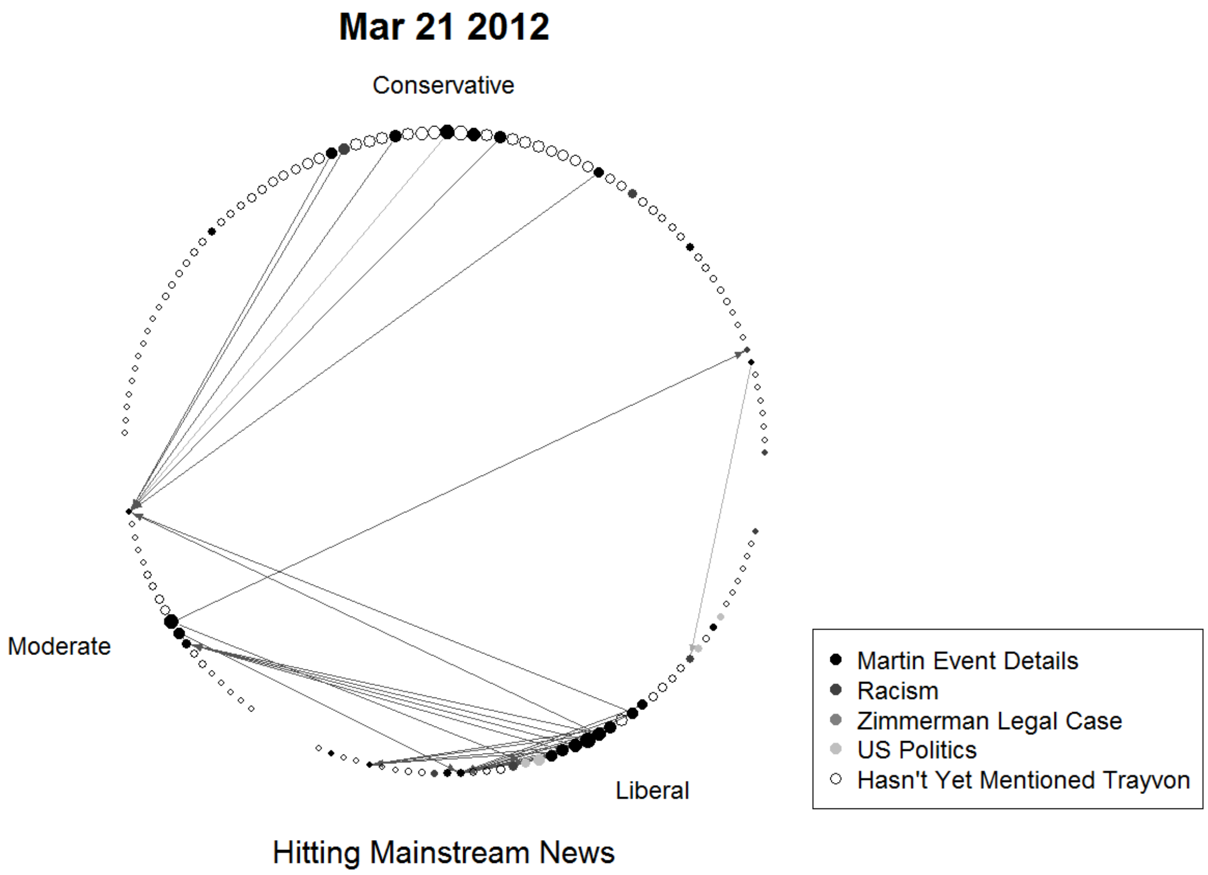
Next we used LDA, which found five topics in the blog posts. To characterize those topics, these are the n-gram tokens that loaded most heavily on each of the topics:

[1]	[2]	[3]	[4]	[5]
money	**obama**	**think**	**georg zimmerman**	**trayvon martin**
log	**presid**	**dont**	**trayvon martin**	**black**
scream	**state**	**comment**	**polic**	**white**
donat	**law**	**would**	**law**	**georg zimmerman**
dave	**american**	**even**	**gun**	**media**
voic	**year**	**peopl**	**case**	**news**
perjuri	**alec**	**like**	**would**	**sharpton**
expert	**govern**	**one**	**self defens**	**racial**
bond	**gun**	**make**	**said**	**year**
paypal	**group**	**get**	**charg**	**hoodi**
owen	**republican**	**thing**	**prosecutor**	**look**
bail	**democrat**	**know**	**evid**	**said**
forens	**war**	**use**	**shot**	**fox**
omara	**nation**	**point**	**state**	**death**
websit	**legisl**	**say**	**shoot**	**race**

LDA topics are interpreted impressionistically:

- Topic 1 seems to focus on the court case (O'Mara is Zimmerman's attorney, Owen is the forensic audiologist, and Zimmerman's disclosure of money raised from supporters on PayPal became a legal issue).
- Topic 2 relates to political aspects of the story.
- Topic 3 reflects the social function of blogging, and absorbs many unspecific words.
- Topic 4 is about the facts in the evolving story.
- Topic 5 is about racism.

The topic weights can be viewed as covariates that enable one to fit a dynamic network model to predict where links will form and how memes get passed among the blogs.



The blog network for the Trayvon Martin discussion when the event was noticed by national media.

3. Network Models

As the movie showed, there is interesting network structure in blogs as well as topic structure. So we want to build a dynamic network model as well.

Their has recently been a lot of work on network models. I'll mostly talk about sociology and statistics, but this omits:

- mathematicians, who developed graph theory, characteristic polynomials, hypergraphs, and tools such as the Laplacian spectrum of a graph;
- biochemists, who pioneered the use of feedback loops and compartmental modeling (e.g., for PK/PD metabolism);
- machine learners, who led the way in fitting dynamic models to very large and very complex data sets.
- physicists, who developed scale-free and preferential attachment models.

Social networks, a major and popular strand of network science, began with work by Georg Simmel in the 1900s and was extended by Jacob Moreno in the 1930s.

Simmel introduced the concept of dyads and triads, emphasizing the fact that human behavior is not well-described in terms of individual choice, but rather needs explanation in terms of relationships among sets of people.

The term **Simmelian ties** refers to three-way relationships, and there is psychological evidence that this is how normative behaviors are enforced. ("God is watching" or "I'll tell Mom".)

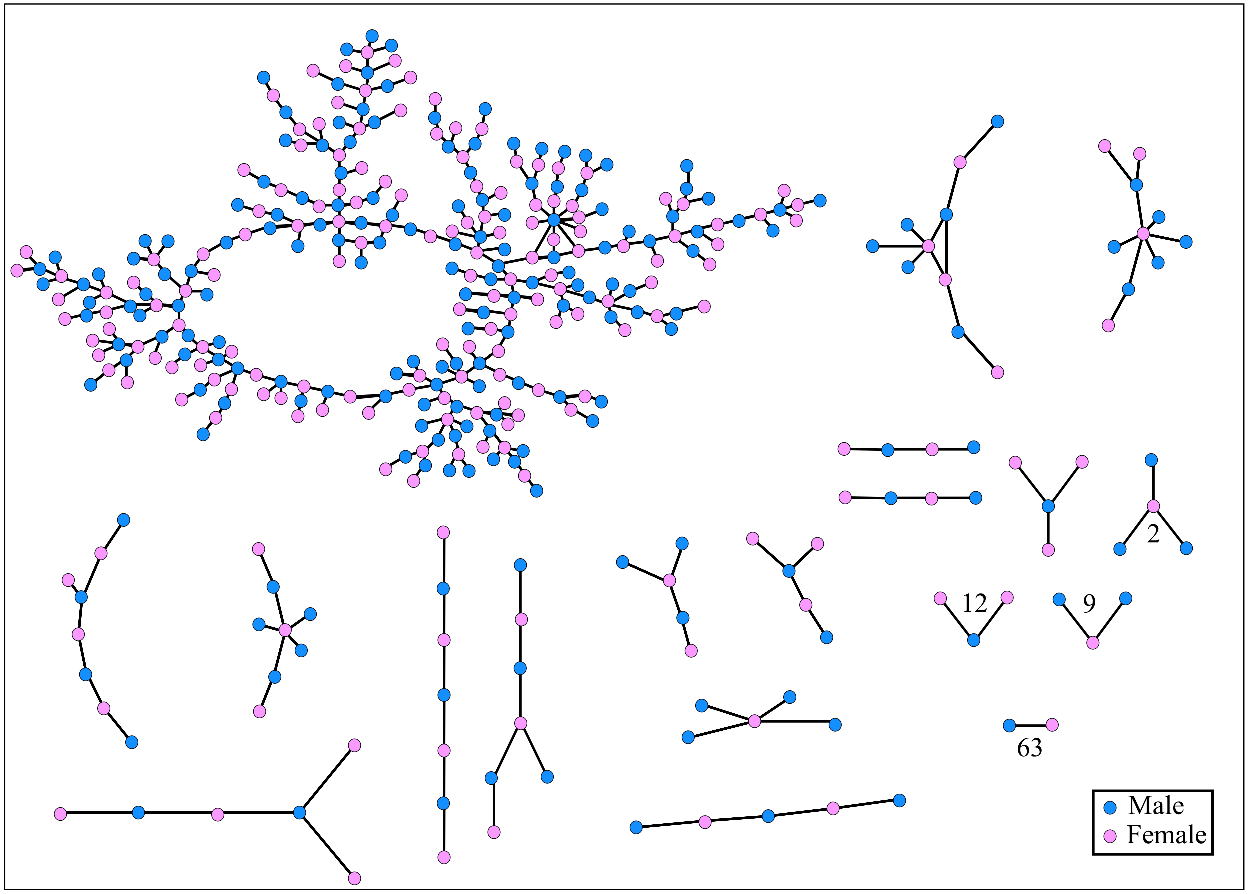
Moreno introduced the sociogram, which is the explicit representation of relationships among agents in terms of nodes and edges in a graph.

This early work sort of petered out, hindered by the difficulty of getting good data and the lack of mathematical sophistication. But sociologists developed some key data sets that have shaped subsequent research:

- Stanley Milgram (1967) contacted random people in Omaha and Wichita, and asked them to relay a package to a target contact person in Boston, with the condition that each stage of the relay had to be someone known on a first-name basis. The analysis supported Karinthy's **six degrees of separation** hypothesis (although it did not properly handle the undelivered packages).
- Samspon (1968) studied social relations (whom they liked, disliked, and sought advice from) among 25 noviates in a monastic community in New England. Measurements were taken at five points in time, which allowed longitudinal insight into an internal conflict that led to the expulsion of four of the aspirants.

Other key datasets include Kapferer's data on relations among female workers in an African cloth factory, and Potterat et al.'s study of sexual relationship networks in Colorado Springs.

An interesting sociology example is a study of “Jefferson High” data on sexual links among 832 high school students reported by Bearman, Moody and Stovel (1995).



.

The Jefferson High data are aggregated over six months, so the STD threat is slightly overemphasized in the graph. But it is clear that there is a large connected component, and many disconnected components.

The researchers found that there were strong tendencies for racial and smoking **homophily**, and for gender heterophily.

The researchers generated random networks that had the same degree counts, the same racial/smoking/gender cross-links, and the simulated networks looked very different (more "clumpy") than the real one.

Then they added an extra condition, to exclude four-cycles. That is, if Bob and Carol are together, and Ted and Alice are together, but later Bob and Alice get together, the new condition says that Ted and Carol cannot get together. (In high school, there is no status gain when two dumpees hook up.)

Simulations with this new constraint looked like the real network.

4. Statistics

The statistical community became involved in 1976, when Holland and Leinhardt introduced the p_1 model. This describes a **directed graph**, in which edges point from one node to another.

The p_1 model says that the probability p_{ij} of an edge from node i to node j satisfies:

$$\ln \frac{p_{ij}}{1 - p_{ij}} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ij}$$

where μ reflects the average connectivity of the network, α_i is the **expansiveness** of node i, β_j represents the **attractiveness** of node j, and γ_{ij} captures **reciprocity**, the increase in p_{ij} that occurs if there is already an edge from node j to node i.

This is just a logistic regression. And all the information is dydadic.

This early model is inadequate. For example, it:

- assumes that all dyads are independent;
- does not account for covariate information;
- treats node characteristics as fixed effects;
- is static rather than dynamic;
- treats only binary relationships (i.e., edges are present or absent; they are not weighted nor multivariate);
- considers only dyads rather than multi-way relationships;
- bidirectional relationships are not realistically described;
- cannot assess goodness-of-fit.

Subsequent work addressed these deficiencies in various ways (cf. Goldenberg, Zheng, Fienberg and Airoldi, 2009).

Nonetheless, this was a revolutionary step forward in the field. It was the first mathematization of the intuitions that sociologists had been discussing.

Frank and Strauss (1986) realized that one could use tools from spatial process theory, and developed the p^* class of models, or the **exponential random graph models** (ERGMs). These models can handle non-dyadic relations.

In the ERGM family, the probabilty of observing a graph g with binary links $\{g_{ij}\}$ is

$$\mathbb{P}[G = g] \propto \exp[\sum_A \eta_A \prod_{g_{ij} \in A} g_{ij}]$$

where A is a graph configuration (e.g., dyads, triads, four-cycles) such that distinct configurations are independent, and η_A is a free parameter that is non-zero if and only if all pairs of variables in A are dependent.

This reduces to the p_1 model when the η_A is non-zero only for edge connections.

As for the p_1 model, there is a straightforward way to include covariate information in an ERGM.

A different kind of extension of the p_1 model is to a latent space model (Hoff, Raftery, and Handcock, 2002). In a latent space model, dyadic relations are explained by two things: observed covariates, and unobserved social distance in a latent space.

The model is

$$\ln \frac{p_{ij}}{1 - p_{ij}} = \mu + \boldsymbol{\alpha}'\boldsymbol{X}_i + \boldsymbol{\beta}'\boldsymbol{X}_j + \boldsymbol{\gamma}'\boldsymbol{X}_{ij} - \|\boldsymbol{z}_i - \boldsymbol{z}_j\| + \epsilon$$

where the first terms are as in the covariate form of the p_1 model, and the last term has a metric $\| \cdot \|$ on the latent space, in which $\boldsymbol{z}_i$ and $\boldsymbol{z}_j$ are the positions of the actors. Since the latter cannot be observed, one has to estimate those (using Markov chain Monte Carlo).

The intuition is that if two actors are distant in the latent space, then they are unlikely to form a link.

Often when one fits a latent space model, it turns out that by examining the estimated positions of the actors, one can infer the axes of the latent space.

5. Conclusions

Network models and topic models are are hot new sciences. With text networks, topics inference can improve network modeling and network models can improve topic discovery.

Key ingredients include logistic regression models to predict link probabilities from covariates, feedback mechanisms, and fits with cluster structure, prescribed degree distributions, and small world features. Latent space ideas have become very popular—they provide automatic cliquing, good visuals, and are often interpretable.

Current research interest is focused on dynamic network data, but our models for such are still weak and largely consist of simple extensions of static models. The two case studies did not get into the modeling details, in part because the methodology is still quite ad hoc.

Movies are a nice way to show network dynamics.