Sina Weibo as a Corpus for Studying Public Opinions

Christine Kuang, Siqi Wu, and Angie Zhu

Department of Statistics, UC Berkeley

May 3, 2012

Outline

Introduction

- Opinions on microblogging and social networking websites
- Sina Weibo 新浪微博 is the largest microblogging website:
  accounted for 65% of China's microblog market as of December 2011
- Study public opinions using Sina Weibo as a corpus for a given topic

Topic

- Internet censorship in China
- Time sensitive
- Processing is topic-dependent
- Hot topic is preferred
- Chosen topic: Han Han 韩寒

Background

- HAN Han 韩寒 (born 23 September 1982) is a Chinese best-selling author, professional rally driver, and wildly popular blogger
- Published his first novel *Triple Gate* 三重门 at age of 17
- High school dropout

Photograph by Tony Law / Redux. Source: http://www.time.com/time/magazine/article/0,9171,1931619,00.html

Background (Cont'd)

- Ghostwriting allegation against Han from January 2012
- FANG Zhouzi 方舟子, a scientific author and anti-fraud crusader, created widespread debate on the Internet
- *Light and Upright 光明与磊落*: photocopied manuscripts set, including his first novel *Triple Gate 三重门*
- Han received a death threat on April 15, 2012

Data Collection

- Topic searching via API:
  - Only the latest results are returned
  - Up to 30 each time
- 22,398 posts collected on April 16 and 17, 2012
- UTF-8 encoding

Characteristics of Chinese Language

- No explicit delimiter
- Ambiguities in phrases
  - Context ambiguition: e.g., 他好吃
  - Word definition ambiguition: e.g., 打
- Out-of-vocabulary words
- No 1-to-1 correspondence between traditional and simplified Chinese

Characteristics of Sina Weibo Posts

1714080953 2012-04-16 10:01:59 //@ 風笑巨石: 那尊神容不得别人质疑? //@ 伯林 2011: 质疑派人士遭到人肉，人身攻击，甚至死亡威胁的时候，从来没有污名化整个挺韩派，也没有引起什么媒体关注，相比之下，韩寒如此炒作，太无良了，挺韩和批韩的双方本不至于如此撕裂

- Multiple forms
- Informal and short
- Reposting: “//@”
- Spams
- Emotion symbols
- Internet slangs
- Topic: “# 话题 #”

Pre-Tagging Processing

1165303315 2012-04-16 09:55:40 《韩寒收到网友死亡威胁》(来自 @ 新浪娱乐) http://t.cn/zOprKap 1165303315 2012-04-16 09:55:40 《Han Han received death threat online》(from @ 新浪娱乐) http://t.cn/zOprKap

- Remove user identification number and time stamp
- Only the reposting user's comment is kept:
  If the resulting string is empty, it will be eliminated as well
- Remove URLs
- Remove duplicates
- 13,070 posts left

Tagging

- Process: tagged 3000 total posts with four categories
- Examples:

  Positive 支持韩寒！ Support Han Han!
  Negative 看到韩寒就恶心。Feel nauseous when I see Han Han.

- Limitations:
  - Subjective responses:
    e.g., “that wasn’t too bad”
  - Uncertain tags
    - Quotes
    - Posts without subjects
    - Posts that just mention opposing author

Pre-Segmentation Processing

- Word segmentation is crucial for our word-based analysis
- Substitute mentioning of topic-related usernames by the corresponding proper nouns
- Remove other mentioned usernames
- Substitute emotional symbols and Internet slangs by the corresponding word surrounded by square brackets

Segmentation

- 汉语词法分析系统 ICTCLAS (Institute of Computing Technology, Chinese Lexical Analysis System) is a well known Chinese word segmentation system
- Chinese word segmentation, lexical tagging, named entity recognition, unknown words detection, and the user-defined dictionary
- Examples from the user-defined dictionary:

  围脖 (wei2 bo2, means "scarf") refers to 微博 (Weibo, wei1 bo2)

  韩少 (韩: Han Han's surname, 少: abbreviation of 少爷, which means "young master of the house") refers to Han Han

Conjunction Rules

[Lee and Renganathan, 2011] suggested that special consideration should be given to

1. Although (part A), (part B).
2. (Part A), but (part B).
3. Although (part A), but (part B).

For each case, only part B will be kept.

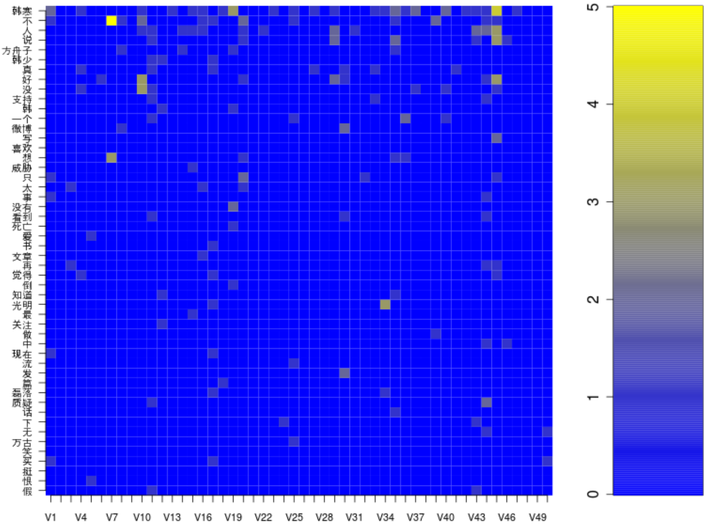
Stop Words and Punctuation Elimination

- Remove prepositions, punctuation marks, English character strings, interjections, modal particles, onomatopoeia, and auxiliary words
- Remove pre-defined stop words and number strings

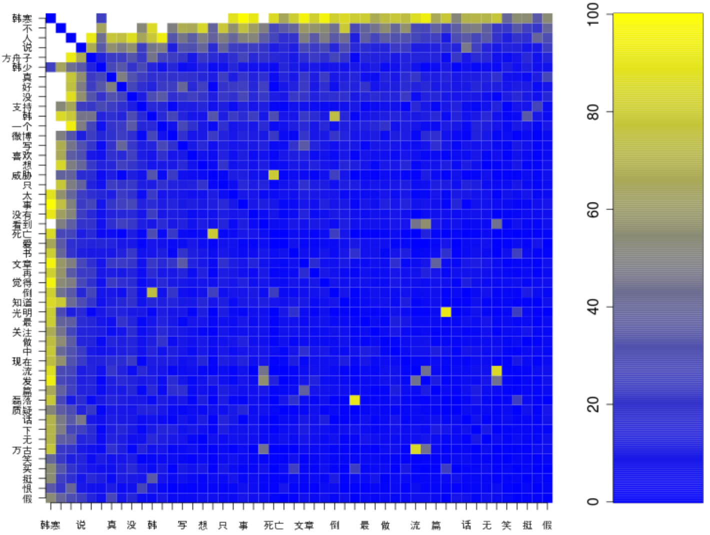
Word Frequency

- Extract the word frequency vector x_i from the i-th post
- Focus on words with overall frequency ≥ 10, resulting in $p = 795$ words
- Construct the word frequency matrix $X = (x_1, ..., x_n)^T$, where $n = 3000$. This will be our design matrix.

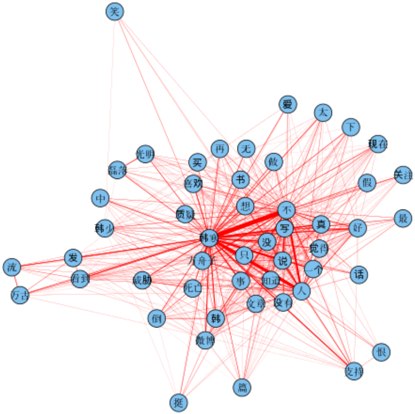
Word Frequency Visualization: Matrix Plot



Co-Occurrance



Co-Occurrance



Sparse Graphical Models

- Goal: study the relation between words by graphical models
- Fact: if $x \in R^p$ follows $N(\mu, \Sigma)$, then for $i \neq j$

$$(x_i \perp\!\!\!\perp x_j) \mid \{x_{\text{all but } (i,j)}\} \textbf{ iff } (\Sigma^{-1})_{ij} = 0$$

- This motivates us to estimate Σ^{-1}.
- Let $x_1, x_2, ...x_n$ be IID $N(\mu, \Sigma)$ data. The joint likelihood of the data is

$$\begin{aligned} & f(x_1, \ldots, x_n | \mu, \Sigma) \\ & = \frac{1}{(2\pi \det(\Sigma))^{n/2}} \exp\left\{ -\frac{1}{2} \sum_{i=1}^{n} (x_i - \mu)^T \Sigma^{-1} (x_i - \mu) \right\}. \end{aligned}$$

Sparse Graphical Models (Cont'd)

- Log-likelihood:

$$l(\mu, \Sigma^{-1}) = -\frac{n}{2} \log \det(\Sigma) - \frac{1}{2} \sum_{i=1}^{n} (x_i - \mu)^T \Sigma^{-1} (x_i - \mu)$$

- Do a maximum likelihood estimation (optimize over μ and $S = \Sigma^{-1}$; easy to see that the MLE for μ is $\bar{x}$):

$$\max_S \left\{ \frac{n}{2} \log \det(S) - \frac{1}{2} \sum_{i=1}^{n} (x_i - \bar{x})^T S (x_i - \bar{x}) \right\}$$

Sparse Graphical Models (Cont'd)

- Trace trick $\sum_{i=1}^{n}(x_i - \bar{x})^T S(x_i - \bar{x}) = \mathbf{Tr}(\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^T S) = n\mathbf{Tr}(\hat{\Sigma}S)$. We end up with:

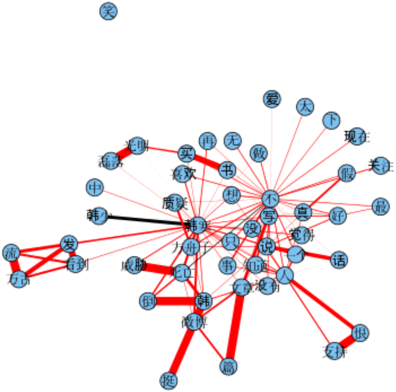
$$\max_S \left\{\log\det(S) - \mathbf{Tr}(\hat{\Sigma}S)\right\}$$

- Fitting a sparse Gaussian graphical model:

$$\max_S \left\{\log\det S - \mathbf{Tr}(\hat{\Sigma}S) - \lambda||S||_1\right\}$$

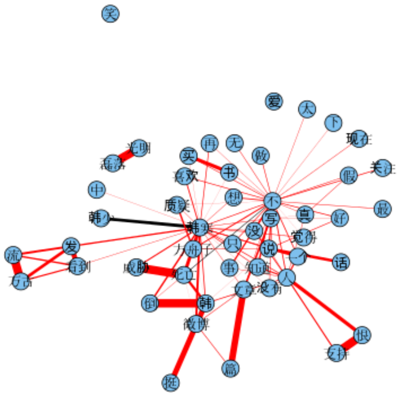
where $||S||_1 = \sum_{i,j}|s_{ij}|$. See, e.g. Banerjee et al. (2007) and Friedman et al. (2007).

Sparse Graphical Models: Results



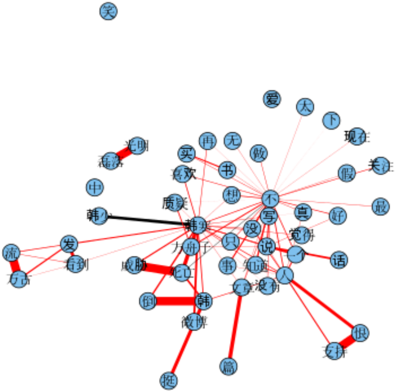
lambda = 0.01

Sparse Graphical Models: Results



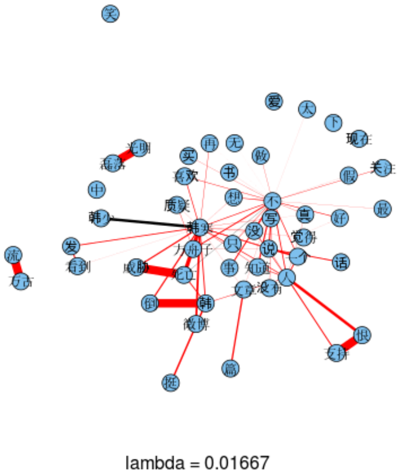
lambda = 0.01222

Sparse Graphical Models: Results

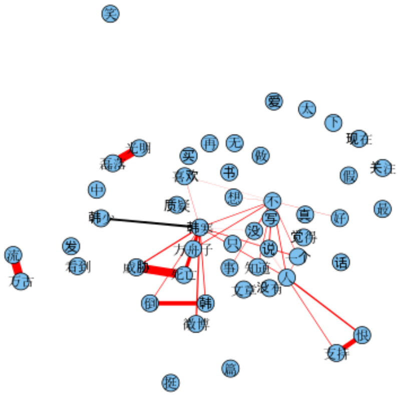


lambda = 0.01444

Sparse Graphical Models: Results

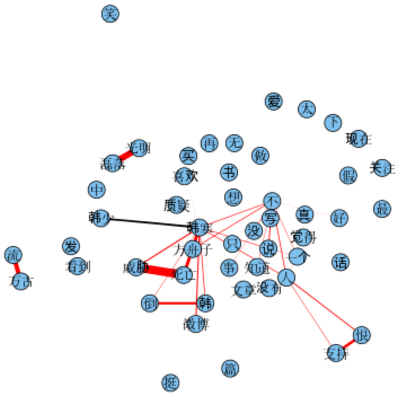


Sparse Graphical Models: Results



lambda = 0.02556

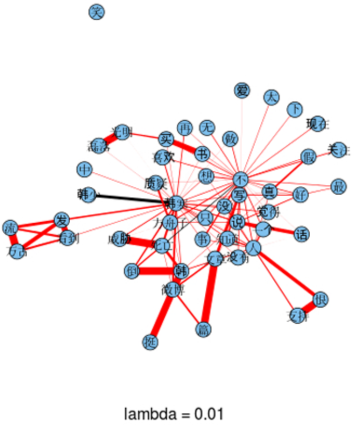
Sparse Graphical Models: Results



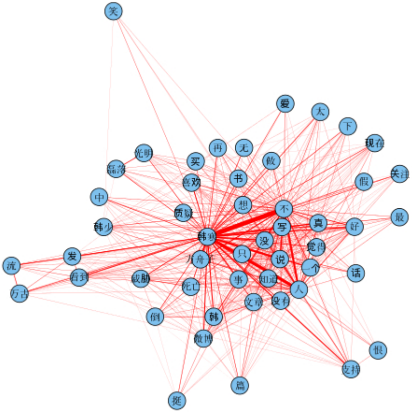
lambda = 0.03

Sparse Graphical Models v.s. Co-Occurence

Sparse Graphical Models



Co-occurence



Classification

- $x_i \in R^p$ be the i-th row of $X \in R^{n \times p}$, where $n = 3000$ and $p = 395$
- y_i: the corresponding category. Assume $y_i \in \{-1, +1\}$, where the $+1$ can have one (and only one) of the following meanings (at a time):
  - positive feeling about Han Han
  - negative feeling about Han Han
  - netural or unidentifiable opinion
  - spam
- Two classification methods: LASSO and l_1-norm SVM.

LASSO

- The Lasso approach (Tibshirani, (1996)):

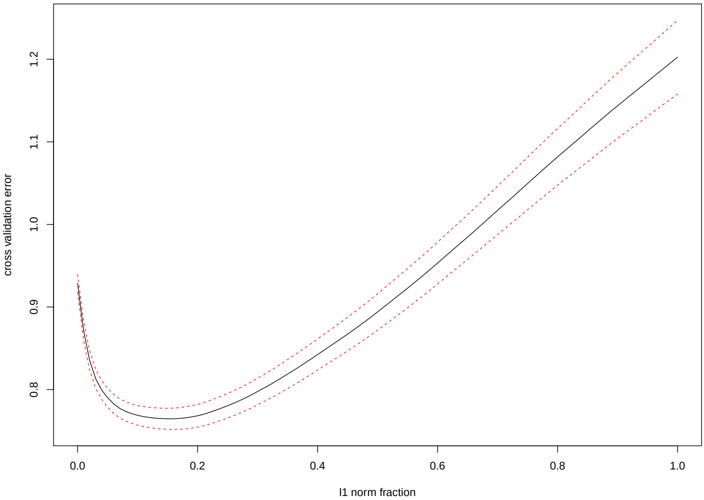
$$\hat{\beta}(\lambda) = \arg\min_{\beta} \frac{1}{2}||y - (\beta_0 + X\beta)||_2^2 + \lambda||\beta||_1$$

- The classifier:

$$\text{class}(x) = \mathbf{sign}(\beta_0 + x^T\beta) \in \{-1, +1\}$$

- Four models for each category for classification
- General overview of method
- General overview of application to data
  - for 4 categories
  - 10 fold CV
  - classification error

Choosing λ: Cross-Validation



LASSO Coefficient interpretation

- The classifier:

$$\text{class}(x) = \textbf{sign}(\beta_0 + x^T\beta) \in \{-1, +1\}$$

- We can look at coefficients β
  - Absolute value: most relevant/predictive words
  - Positive: more likely to classify the post in $+1$ category (all other covariates being fixed)
  - Negative: more likely to be in -1 category

Positive v.s. Nonpositive Classification Result

- $y_i \in \{-1, +1\}$;
- $+1$: positive opinion (about Han Han);
- -1: non-positive opinion, including negative, neutral and spam.

Word	Absolute Coef.	Word	Positive Coef.	Word	Negative Coef.
加油 (keep going)	0.820	加油 (keep going)	0.820	样子 (manner)	-0.396
韩少 (Master Han)	0.644	韩少 (Master Han)	0.644	恋 (love)	-0.344
成熟 (mature)	0.546	成熟 (mature)	0.546	发表 (announce)	-0.336
顶 (support)	0.533	顶 (support)	0.533	道理 (rational)	-0.336
宽容 (tolerant)	0.518	宽容 (tolerant)	0.518	利益 (benefit)	-0.335

LASSO word images for the positive v.s. nonpositive classification.

Negative v.s. Nonnegative Classification Result

- $+1$: negative opinion;
- -1: non-negative opinion, including positive, neutral and spam.

Word	Absolute Coef.	Word	Positive Coef.	Word	Negative Coef.
讨厌 (hate)	0.481	讨厌 (hate)	0.481	支持 (support)	-0.008
无耻 (shameless)	0.412	无耻 (shameless)	0.412	-	-
恶心 (disgusting)	0.395	恶心 (disgusting)	0.395	-	-
骗子 (liar)	0.380	骗子 (liar)	0.380	-	-
扁 (beat up)	0.353	扁 (beat up)	0.353 -	- -	-

LASSO

word images for the negative v.s. nonnegative classification.

Standard Support Vector Machine

- Again, linear decision function $f(x) = \beta_0 + \beta x$;
- The classifier class$(x) = \textbf{sign}(f(x))$.
- The support vector machine (SVM) (see, e.g. Hastie et al 2001):

$$\min_{\beta_0, \beta} \sum_{i=1}^{n} (1 - y_i f(x_i))_+ + \frac{\lambda}{2} ||\beta||_2^2,$$

where $z_+ = \max(0, z)$.

l_1-Norm Support Vector Machine

- Replacing the l_2-norm by l_1-norm yields the sparse SVM (Zhu et al 2003):

$$\min_{\beta_0,\beta}\left\{\sum_{i=1}^{n}(1-y_i f(x_i))_+ + \lambda\|\beta\|_1\right\}.$$

- Computation: use the matlab **lpsvm** package by Fung and Mangasarian (2002)

Positive v.s. Nonpositive Classification Result

- $+1$: positive opinion;
- -1: non-positive opinion, including negative, neutral and spam.
- Cross validation result:
  - training sample misclassification rate: 16.9%
  - testing sample misclassification rate: 28.2%

Word	Absolute Coef.	Word	Positive Coef.	Word	Negative Coef.
加油 (keep going)	2.340	加油 (keep going)	2.340	铁证 (clear evidence)	2.305
铁证 (clear evidence)	2.305	家人 (family)	2.269	接受 (accept)	2.061
家人 (family)	2.269	韩少 (Master Han)	1.969	媒体 (media)	1.907
接受 (accept)	2.061	成熟 (mature)	1.806	默默 (quietly)	1.883
韩少 (Master Han)	1.969	顶 (support)	1.803	四娘 (GUO Jingming)	1.762

l_1-SVM word images for the positive v.s. positive classification.

Negative v.s. Nonnegative Classification Result

- $+1$: negative opinion;
- -1: non-negative opinion, including positive, neutral and spam.
- Cross validation result:
  - training sample misclassification rate: 6.4%
  - testing sample misclassification rate: 11.5%

Word	Absolute Coef.	Word	Positive Coef.	Word	Negative Coef.
扁 (beat up)	1.777	扁 (beat up)	1.777	脑子 (mind)	1.447
苦肉计 (the use of self-injury to win somebody's confidence)	1.708	苦肉计 (the use of self-injury to win somebody's confidence)	1.708	彻底 (completely)	1.290
恶心 (disgusting)	1.527	恶心 (disgusting)	1.527	送给 (give)	1.221
脑子 (asdf)	1.447	骗子 (liar)	1.301	感觉 (feel)	1.109
骗子 (liar)	1.301	公开 (open)	1.220	热点 (hot interest)	1.101

l_1-SVM word images for the negative v.s. nonnegative classification.

l_1-Norm SVM v.s. LASSO

Positive v.s. Nonpositive Classification Results
Positive coefficients

l_1-Norm SVM		LASSO	
Word	Postive Coef.	Word	Positive Coef.
加油 (keep going)	2.340	加油 (keep going)	0.820
家人 (family)	2.269	韩少 (Master Han)	0.644
韩少 (Master Han)	1.969	成熟 (mature)	0.546
成熟 (mature)	1.806	顶 (support)	0.533
顶 (support)	1.803	宽容 (tolerant)	0.518

Further Work

- Comparison with maximum entropy approach
- Graphical model to track reposting
- Statistical models for identifying Internet slangs
- Sampling from large graphs