Deepening Our Understanding of Social Media via Data Mining

Huan Liu

Joint work with

Shamanth Kumar

Ali Abbasi

Reza Zafarani

Fred Morstatter

Jiliang Tang

Social Media Mining by Cambridge University Press

Social Media Mining

Home | Book | Errata | Slides | Table of Contents | Tutorials

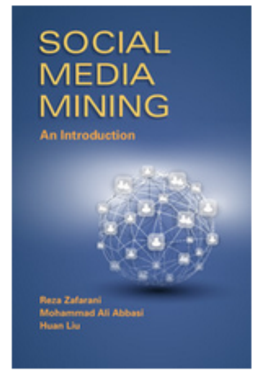
Social Media Mining

An Introduction

A Textbook by Cambridge University Press

Reza Zafarani — *Arizona State University*
Mohammad Ali Abbasi — *Arizona State University*
Huan Liu — *Arizona State University*

Download



CAMBRIDGE UNIVERSITY PRESS

amazon.com

BARNES & NOBLE BOOKSELLERS

eBooks.com

The growth of social media over the last decade has revolutionized the way individuals interact and industries conduct business. Individuals produce data at an unprecedented rate by interacting, sharing, and consuming content through social media. Understanding and processing this new type of data to glean actionable patterns presents challenges and opportunities for interdisciplinary research, novel algorithms, and tool development. Social Media Mining integrates social media, social network analysis, and data mining to provide a convenient and coherent platform for students, practitioners, researchers, and project managers to understand the basics and potentials of social media mining. It introduces the unique problems arising from social media data and presents fundamental concepts, emerging issues, and effective algorithms for network analysis and data mining. Suitable for use in advanced undergraduate and beginning graduate courses as well as professional short courses, the text contains exercises of different degrees of difficulty that improve understanding and help apply concepts, principles, and methods in various scenarios of social media mining.

http://dmml.asu.edu/smm/

Traditional Media and Data

Broadcast Media
One-to-Many

Communication Media
One-to-One

Traditional Data

Social Media: Many-to-Many

- Everyone can be a media outlet or producer
- Disappearing communication barrier
- Distinct characteristics
  - User generated content: Massive, dynamic, extensive, instant, and noisy
  - Rich user interactions: Linked data
  - Collaborative environment, and wisdom of the crowd
  - Many small groups (the long tail phenomenon)
  - Attention is expensive

Unique Features of Social Media

- Novel phenomena observed from people's *interactions* in social media
- Unprecedented opportunities for *interdisciplinary and collaborative* research
  - How to use social media to study human behavior?
    - It's rich, noisy, free-form, and definitely BIG
  - With so much data, how can we **make sense** of it?
    - Putting "bricks" into a useful (meaningful) "edifice"
    - Developing new methods/tools for social media mining

Some Challenges in Mining Social Media

- A Big-Data Paradox
  - How big is the big social media data?
- Studying Distrust in Social Media
  - Is distrust simply the negation of trust? Where to find distrust information with “one-way” relations?
- Sampling Bias
  - Often we get a small sample of (still big) data. How can we ensure if the data can lead to credible findings?
- Noise-Removal Fallacy
  - How do we remove noise without losing too much?

A Big-Data Paradox

- Collectively, social media data is indeed big
- For an individual, there is little data on a site
  - How much activity data do we generate daily?
  - How many posts did we post this week?
  - How many friends do we have?
- Often, we use different social media services for varied purposes
  - Facebook, Twitter, Instagram, YouTube, …
- “Big” social media data often may not be big
  - Searching for more data with limited data

An Example

Little data about an individual

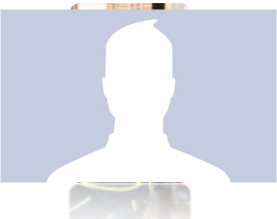
Many social media sites

Partial Information

Complementary Information

Better User Profiles

Reza Zafarani



	LinkedIn	Twitter
Age	N/A	N/A
Location	Phoenix Area	Tempe, AZ
Education	ASU (2014)	ASU

Connectivity is not available

Consistency in Information Availability

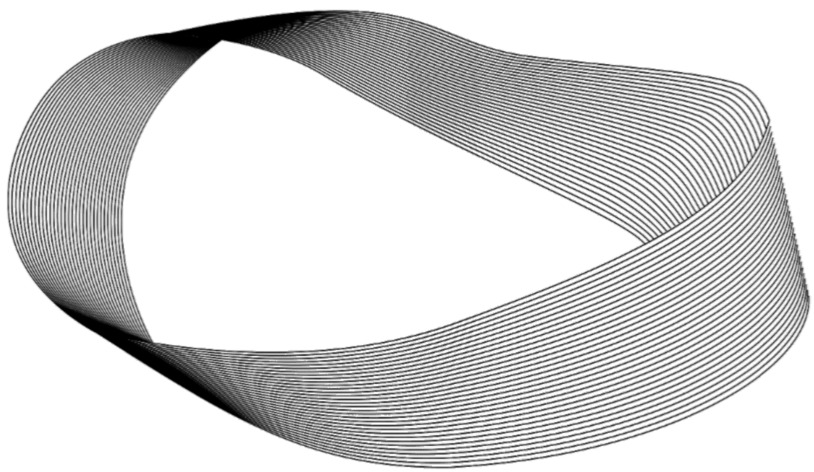
Can we connect individuals across sites?

Searching for More Data with Limited Data

- Each social media site can have varied amount of user information
- What is guaranteed to exist for the joint set of these sites?
  - Usernames
- A user's usernames on different sites can be different
- We set out to verify that the information provided across sites belong to the same individual

Our Behavior Generates Information Redundancy

- Information shared across sites provides a behavioral fingerprint

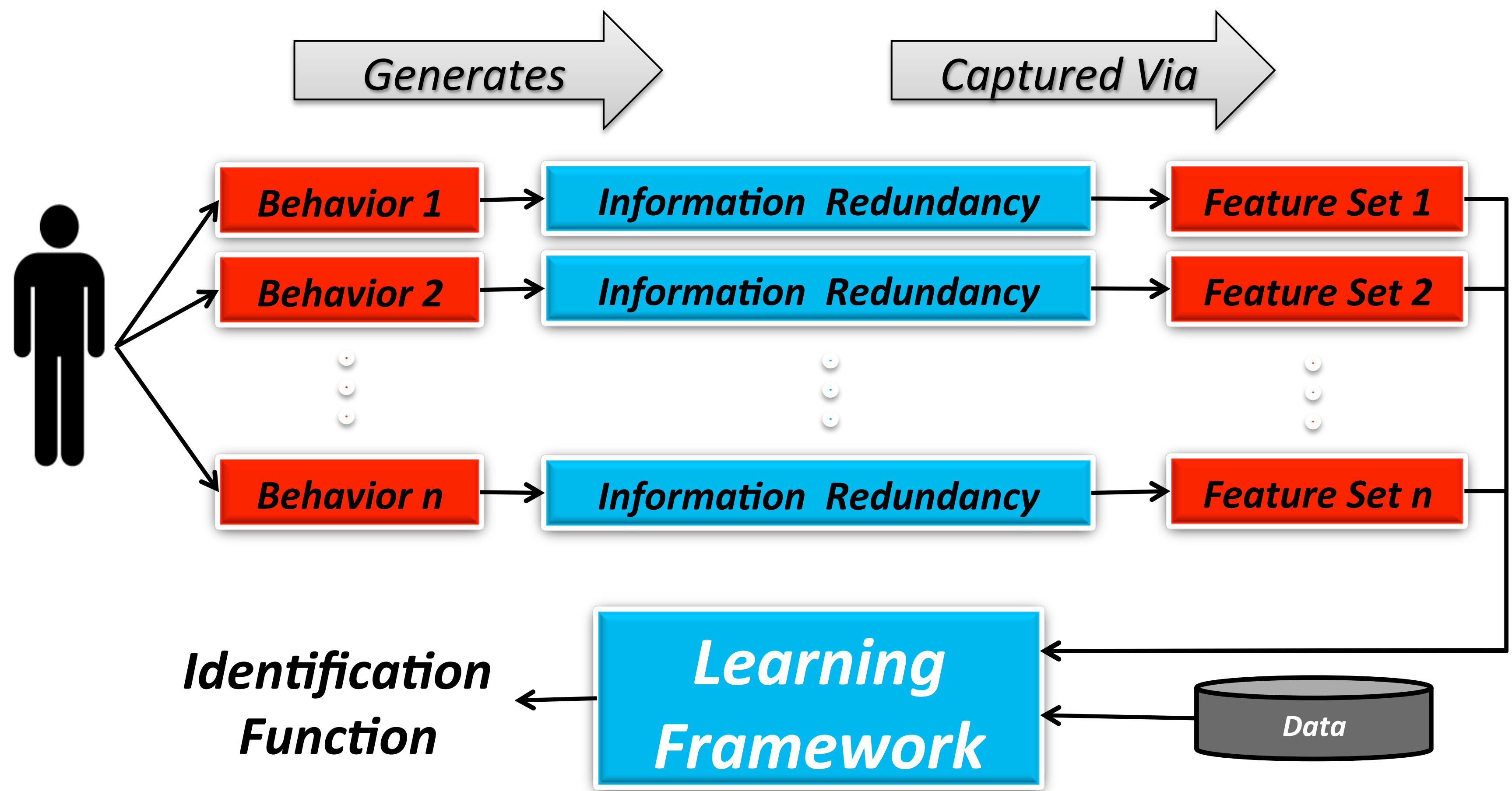


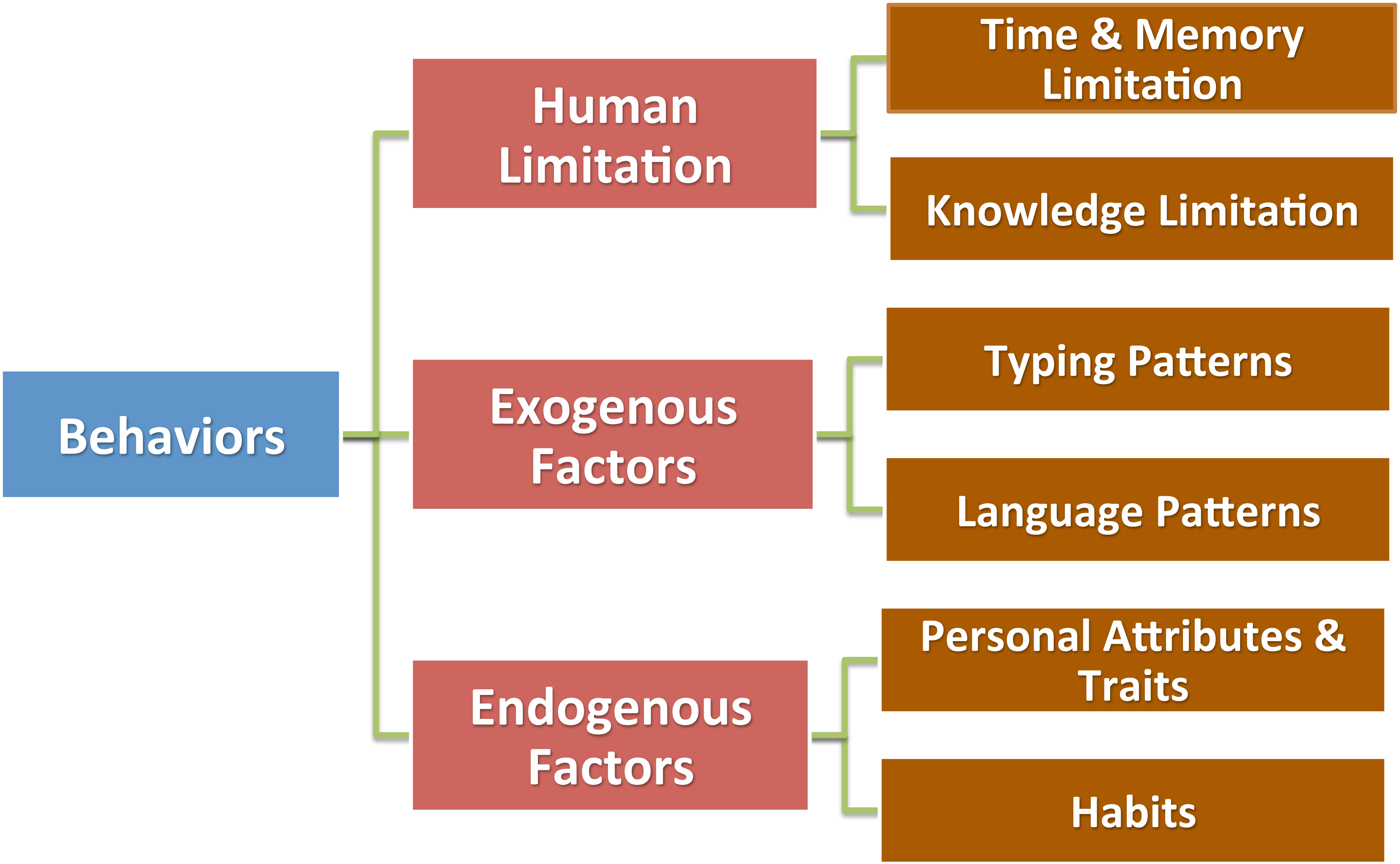
MOBIUS

- **Behavioral Modeling**
- **Minimum Information**

MOdeling **B**ehavior for **I**dentifying **U**sers across **S**ites

Starting with Minimum Information of a User



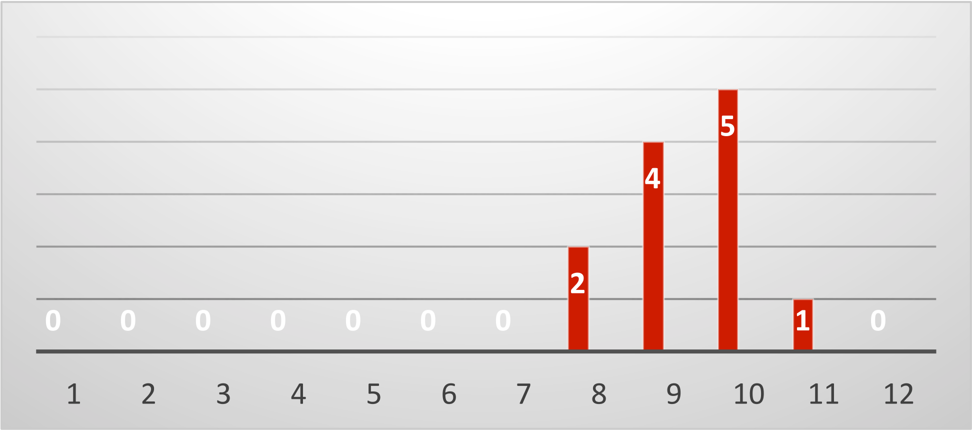

Behaviors
Human Limitation
Exogenous Factors
Endogenous Factors
Time & Memory Limitation
Knowledge Limitation
Typing Patterns
Language Patterns
Personal Attributes & Traits
Habits

Time and Memory Limitation

Using Same Usernames

59% of individuals use the same username

Username Length Likelihood



Knowledge Limitation

Limited Vocabulary

Identifying individuals by their vocabulary size

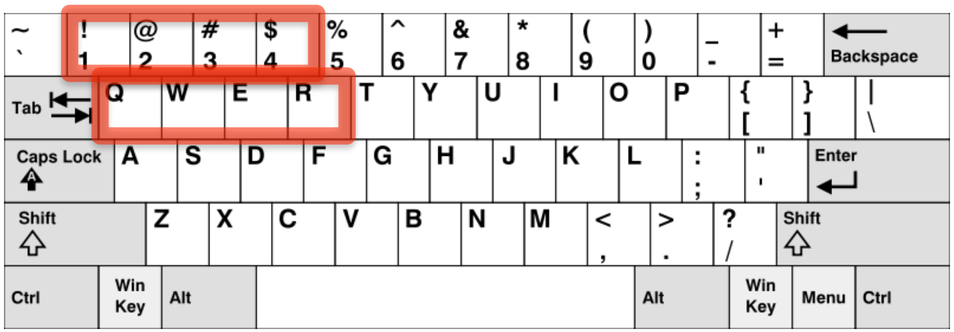
Limited Alphabet

Alphabet Size is correlated to language:

शमंत कुमार -> Shamanth Kumar

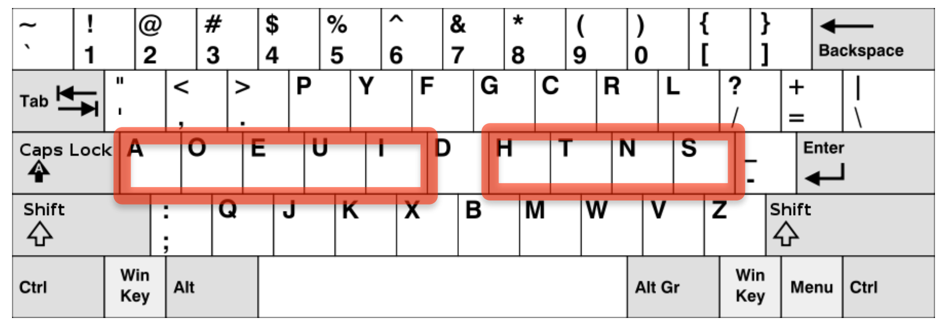
Typing Patterns

QWER1234



QWERTY Keyboard
Variants: AZERTY, QWERTZ

AOEUISNTH



DVORAK Keyboard

Keyboard type impacts your usernames

We compute features that capture typing patterns:
the distance you travel for typing the username,
the number of times you change hands when typing it, etc.

Habits - old habits die hard

Modifying Previous Usernames

Adding Prefixes/Suffixes, Abbreviating, Swapping or Adding/Removing Characters

Creating Similar Usernames

Nametag and Gateman

Username Observation Likelihood

Usernames come from a language model

Obtaining Features from Usernames

For each username:

414 Features

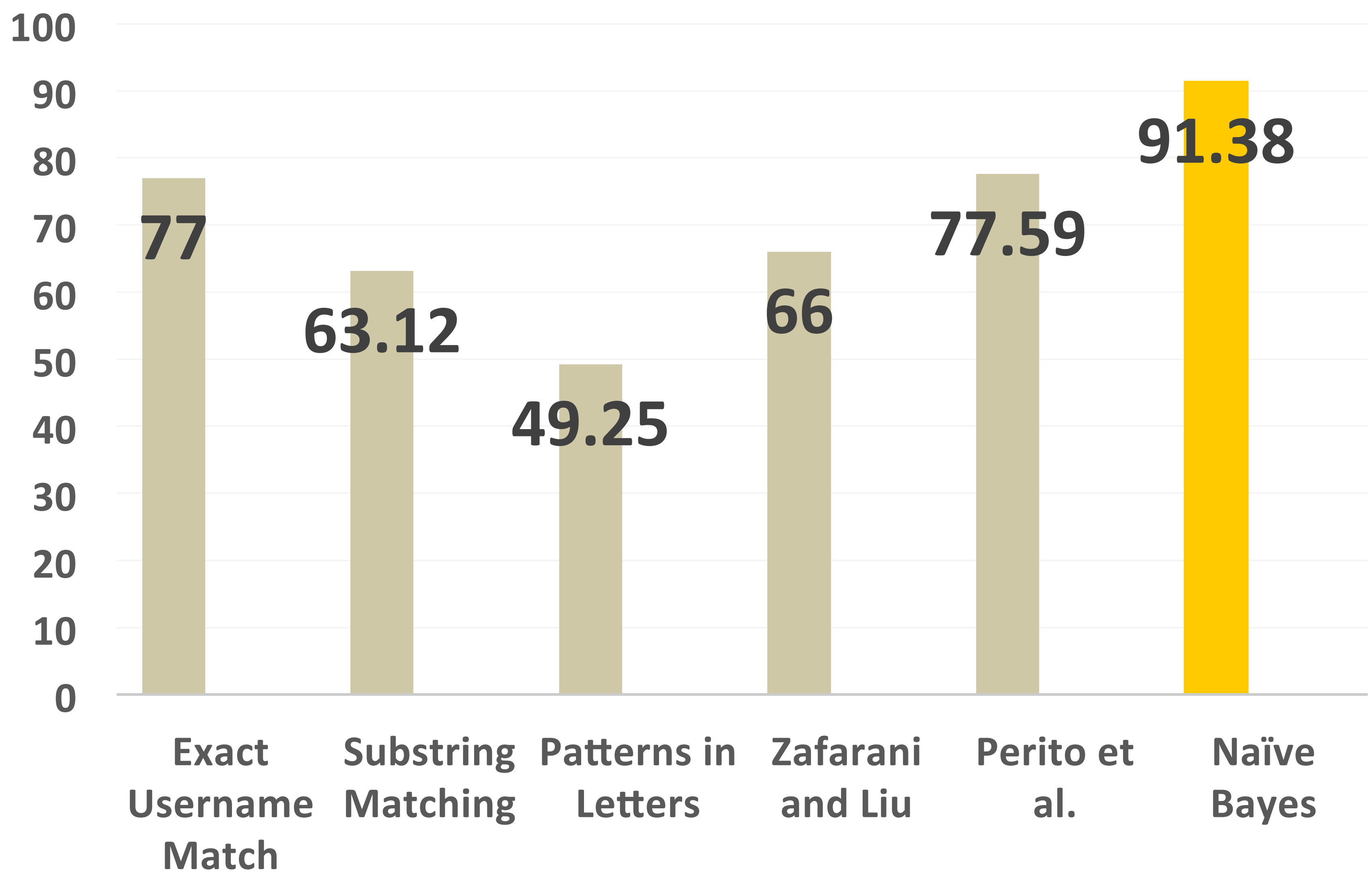
Similar Previous Methods:

1) Zafarani and Liu, 2009
2) Perito et al., 2011

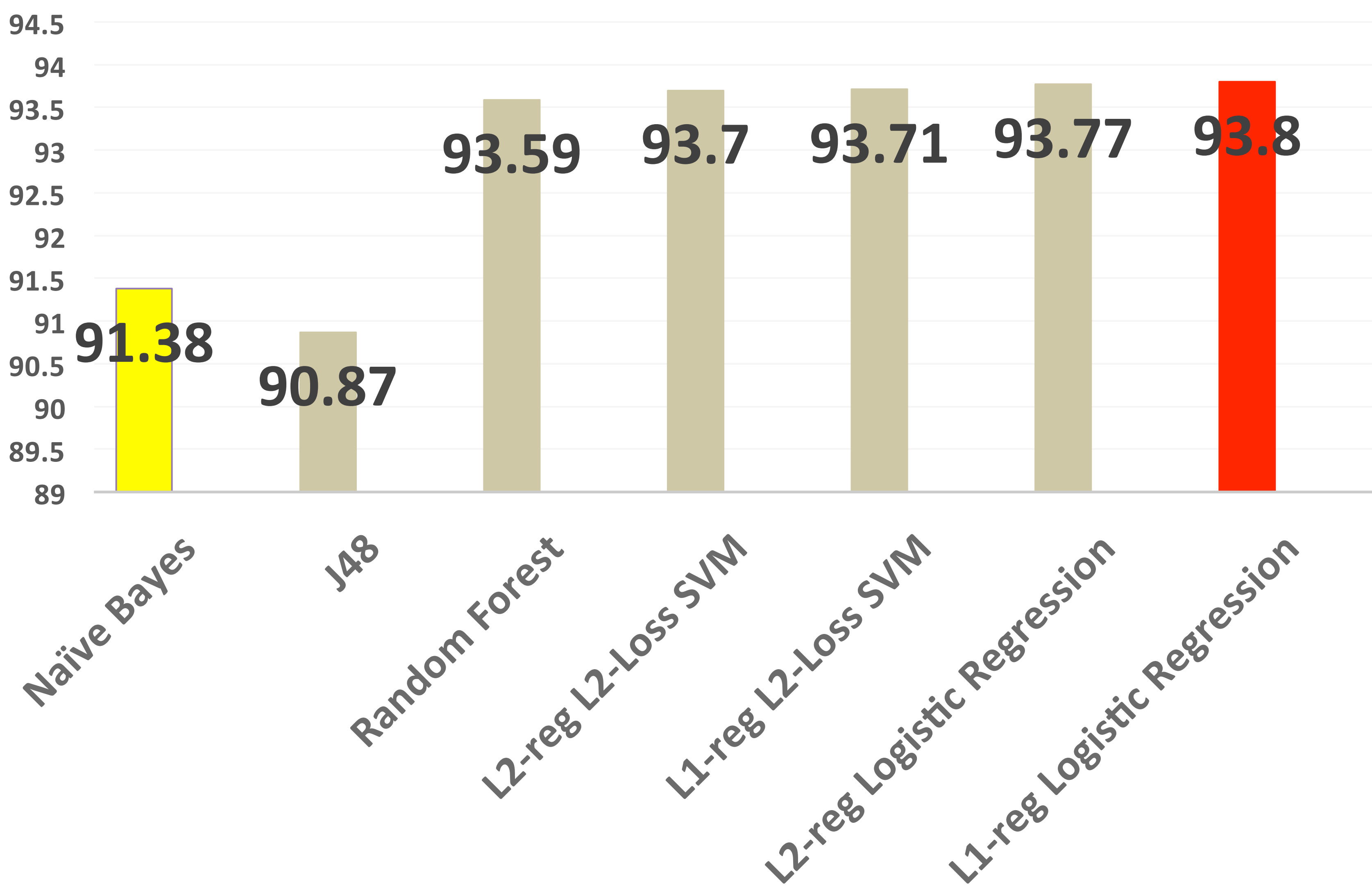
Baselines:

1) Exact Username Match
2) Substring Match
3) Patterns in Letters

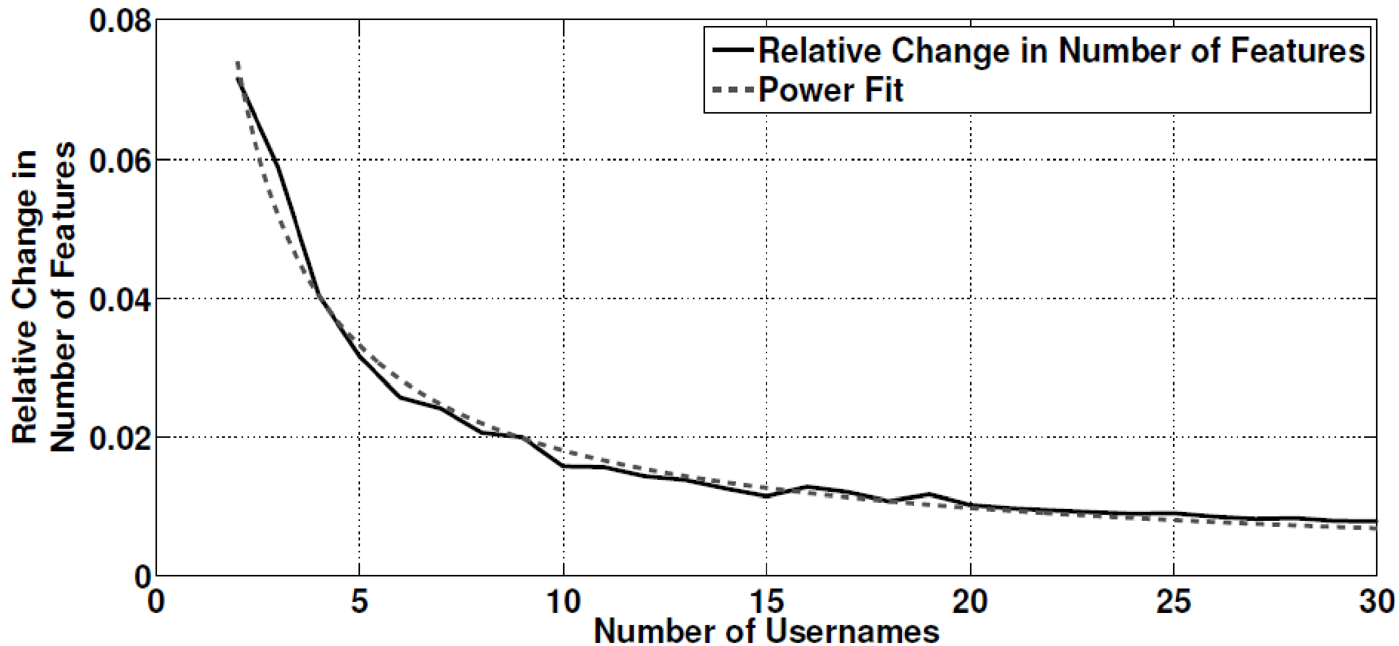
MOBIUS Performance



Choice of Learning Algorithm



Diminishing Returns for Adding More Usernames



Summary

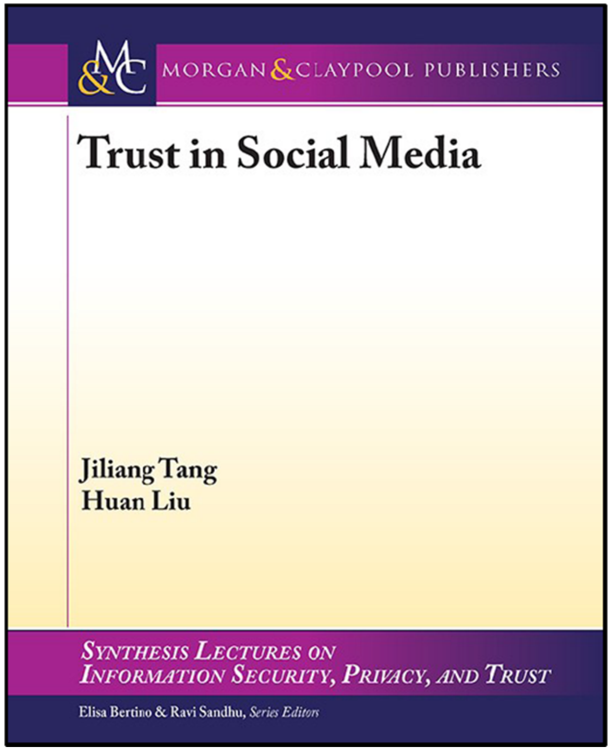
- Many a time, big data may not be sufficiently big for a data mining task
- Gathering more data is often necessary for effective data mining
- Social media data provides unique opportunities such as numerous sites and abundant user-generated content
- Traditionally available data can be equally tapped for making data "thicker"

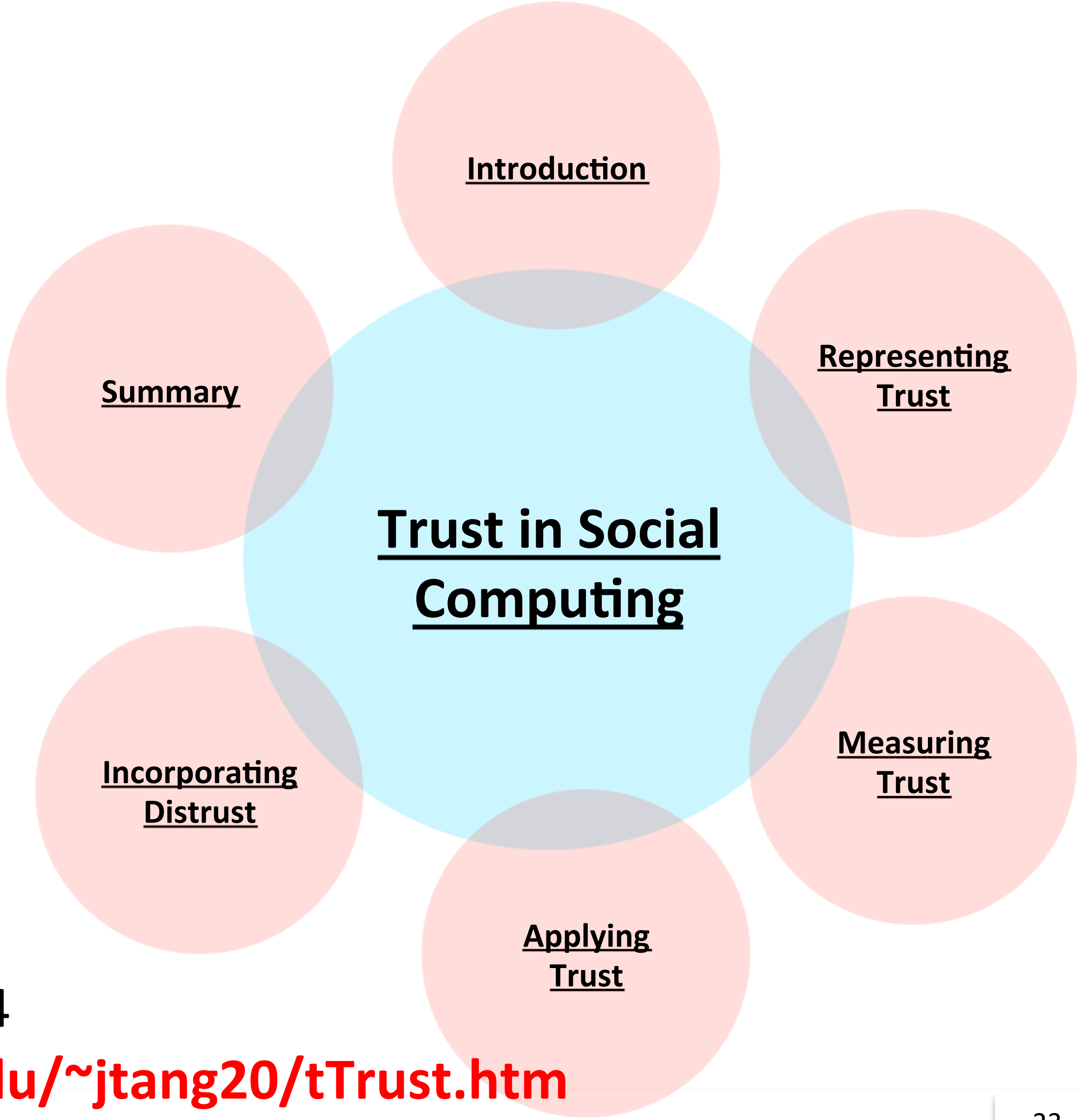
Reza Zafarani and Huan Liu. ``Connecting Users across Social Media Sites: A Behavioral-Modeling Approach", SIGKDD, 2013.

Some Challenges in Mining Social Media

- A Big-Data Paradox
- Studying Distrust in Social Media
- Sampling Bias
- Noise-Removal Fallacy

Studying Distrust in Social Media





WWW2014 Tutorial on
Trust in Social Computing
Seoul, South Korea. 4/7/14
http://www.public.asu.edu/~jtang20/tTrust.htm

Distrust in Social Sciences

- Distrust can be as important as trust

- Both trust and distrust help a decision maker reduce the uncertainty and vulnerability associated with decision consequences

- Distrust may play an equally important, if not more, critical role as trust in consumer decisions

Understandings of Distrust from Social Sciences

- Distrust is the negation of trust
  - Low trust is equivalent to high distrust
  - The absence of distrust means high trust
  - Lack of the studying of distrust matters little
- Distrust is a new dimension of trust
  - Trust and distrust are two separate concepts
  - Trust and distrust can co-exist
  - A study ignoring distrust would yield an incomplete estimate of the effect of trust

Distrust in Social Media

- Distrust is rarely studied in social media
- Challenge 1: Lack of computational understanding of distrust with social media data
  - Social media data is based on passive observations
  - Lack of some information social sciences use to study distrust
- Challenge 2: Distrust information is usually not publicly available
  - Trust is a desired property while distrust is an unwanted one for an online social community

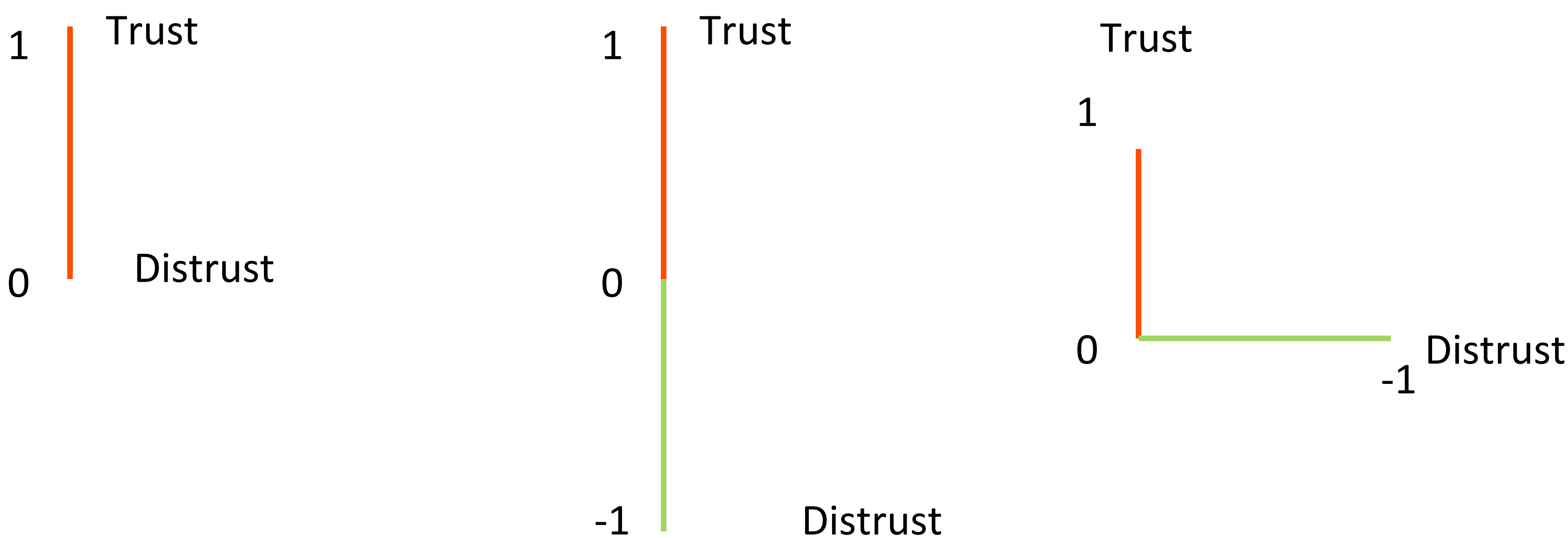
Computational Understanding of Distrust

- Design computational tasks to help understand distrust with passively observed social media data
  - Task 1: Is distrust the negation of trust?
    - If distrust is the negation of trust, distrust should be predictable from only trust
  - Task 2: Can we predict trust better with distrust?
    - If distrust is a new dimension of trust, distrust should have added value on trust and can improve trust prediction
- The first step to understand distrust is to make distrust computable by incorporating distrust in trust models

Distrust in Trust Representations

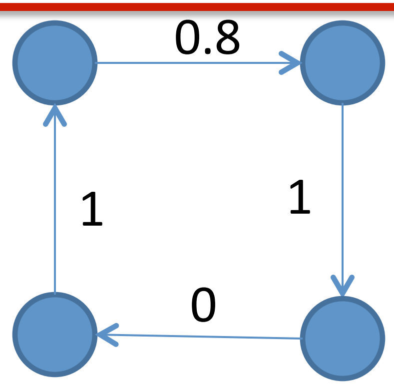
There are three major ways to incorporate distrust in trust representation

- Considering low trust as distrust
- Adding signs to trust values
- Adding a dimension in trust representations

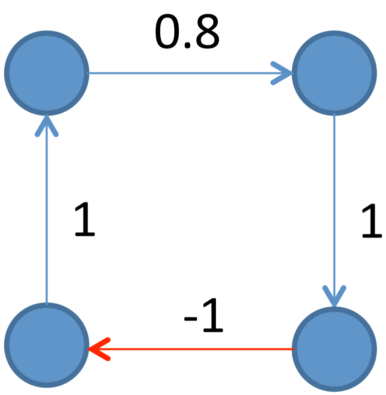


An Illustration of Distrust in Trust Representations

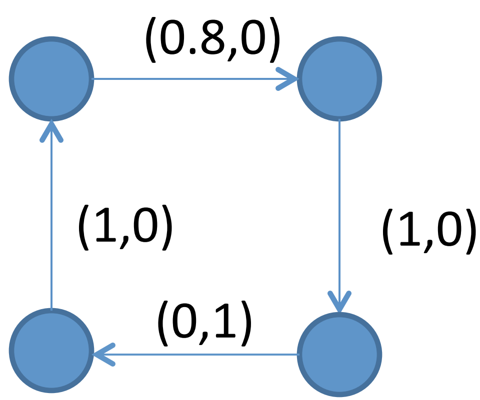
- Considering low trust as distrust
  - Weighted unsigned network



- Extending negative values
  - Weighted signed network



- Adding another dimension
  - Two-dimensional unsigned network



Task 1: Is Distrust the Negation of Trust?

- If distrust is the negation of trust, low trust is equivalent to distrust and distrust should be predictable from trust

IF	Distrust	≡	Low Trust
THEN	Predicting Distrust	≡	Predicting Low Trust

- Given the transitivity of trust, we resort to trust prediction algorithms to compute trust scores for pairs of users in the same trust network

Evaluation of Task 1

- The performance of using low trust to predict distrust is consistently worse than randomly guessing
- Task 1 fails to predict distrust with only trust; and distrust is not the negation of trust

x (%)	dTP ($\times 10^{-5}$)	dMF($\times 10^{-5}$)	dTP-MF($\times 10^{-5}$)	Random($\times 10^{-5}$)
50	4.8941	4.8941	4.8941	5.6824
55	5.6236	5.6236	5.6236	8.1182
60	7.1885	7.1885	7.1885	15.814
65	11.985	11.985	11.985	19.717
70	13.532	13.532	13.532	18.826
80	10.844	10.844	10.844	16.266
90	12.720	12.720	12.720	25.457
100	14.237	14.237	14.237	29.904

dTP: It uses trust propagation to calculate trust scores for pairs of users

dMF: It uses the matrix factorization based predictor to compute trust scores for pairs of users

dTP-MF: It is the combination of dTP and dMF using OR

Task 2: Can we predict Trust better with Distrust

- If distrust is not the negation of trust, distrust should provide additional information about users, and could have added value beyond trust

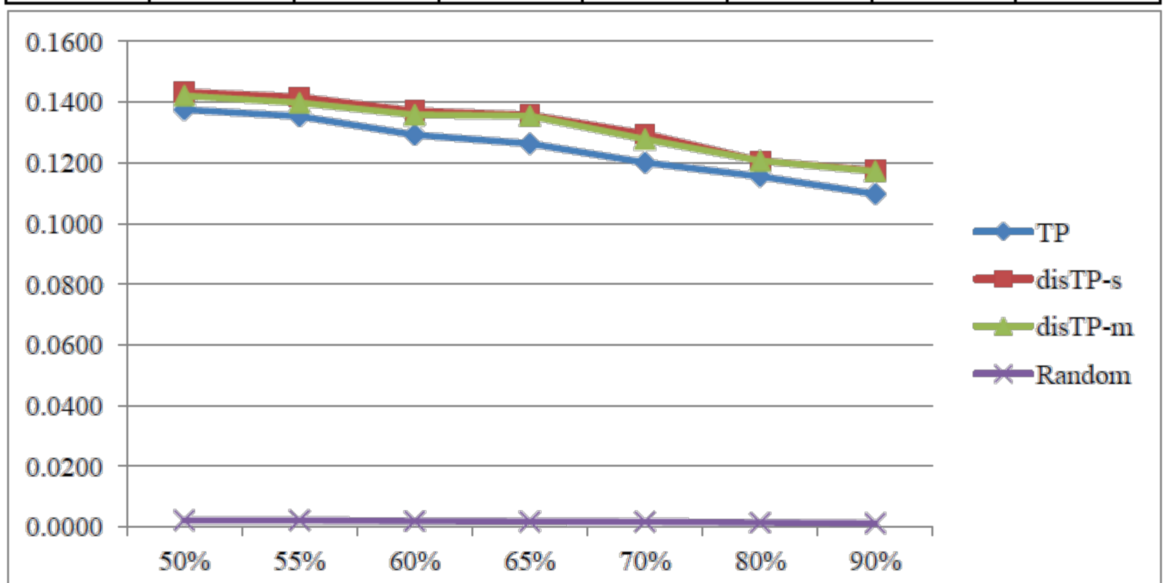
- We seek answer to whether using both trust and distrust information can help achieve better performance than using only trust information

- We can add distrust propagation in trust propagation to incorporate distrust

Evaluation of Trust and Distrust Propagation

- Incorporating distrust propagation into trust propagation can improve the performance of trust measurement
- One step distrust propagation usually outperforms multiple step distrust propagation

	50%	55%	60%	65%	70%	80%	90%
TP	**0.1376**	0.1354	0.1293	0.1264	0.1201	0.1156	0.1098
disTP-s	**0.1435**	0.1418	0.1372	0.1359	0.1296	0.1207	0.1176
disTP-m	**0.1422**	0.1398	0.1359	0.1355	0.1279	0.1207	0.1173
Random	**0.0023**	0.0023	0.0020	0.0019	0.0018	0.0015	0.0013



Some Challenges in Mining Social Media

- A Big-Data Paradox
- Studying Distrust in Social Media
- Sampling Bias
- Noise-Removal Fallacy

Sampling Bias in Social Media Data

- Twitter provides two main outlets for researchers to access tweets in real time:
  - Streaming API (~1% of all public tweets, <u>free</u>)
  - Firehose (100% of all public tweets, <u>costly</u>)
- Streaming API data is often used to by researchers to validate hypotheses.
- How ***well*** does the sampled Streaming API data measure the true activity on Twitter?

Facets of Twitter Data

- Compare the data along different facets
- Selected facets commonly used in social media mining:
  - Top Hashtags
  - **Topic Extraction**
  - Network Measures
  - Geographic Distributions

Preliminary Results

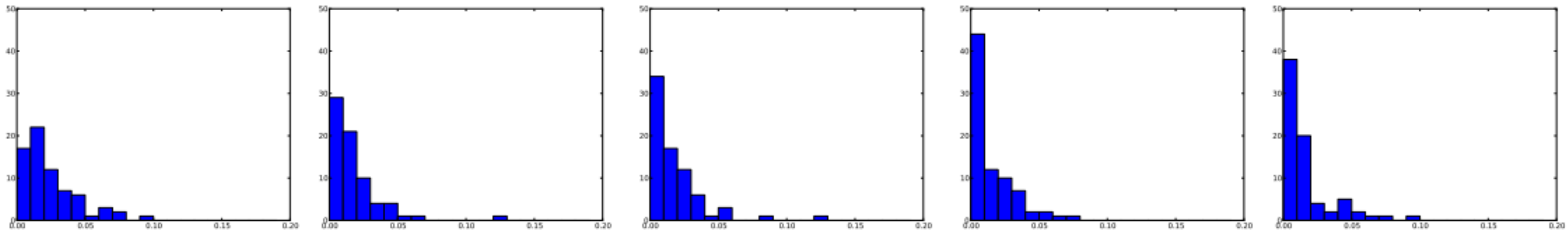
Top Hashtags	Topic Extraction
• No clear correlation between Streaming and Firehose data.	• Topics are close to those found in the Firehose.

Network Measures	Geographic Distributions
• Found ~50% of the top tweeters by different centrality measures. • Graph-level measures give similar results between the two datasets.	• Streaming data gets >90% of the geotagged tweets. • Consequently, the distribution of tweets by continent is very similar.

How are These Results?

- Accuracy of streaming API can vary with analysis to be performed
- These results are about single cases of streaming API
- Are these findings significant, or just an artifact of random sampling?
- How do we verify that our results indicate sampling bias or not?

Histogram of JS Distances in Topic Comparison

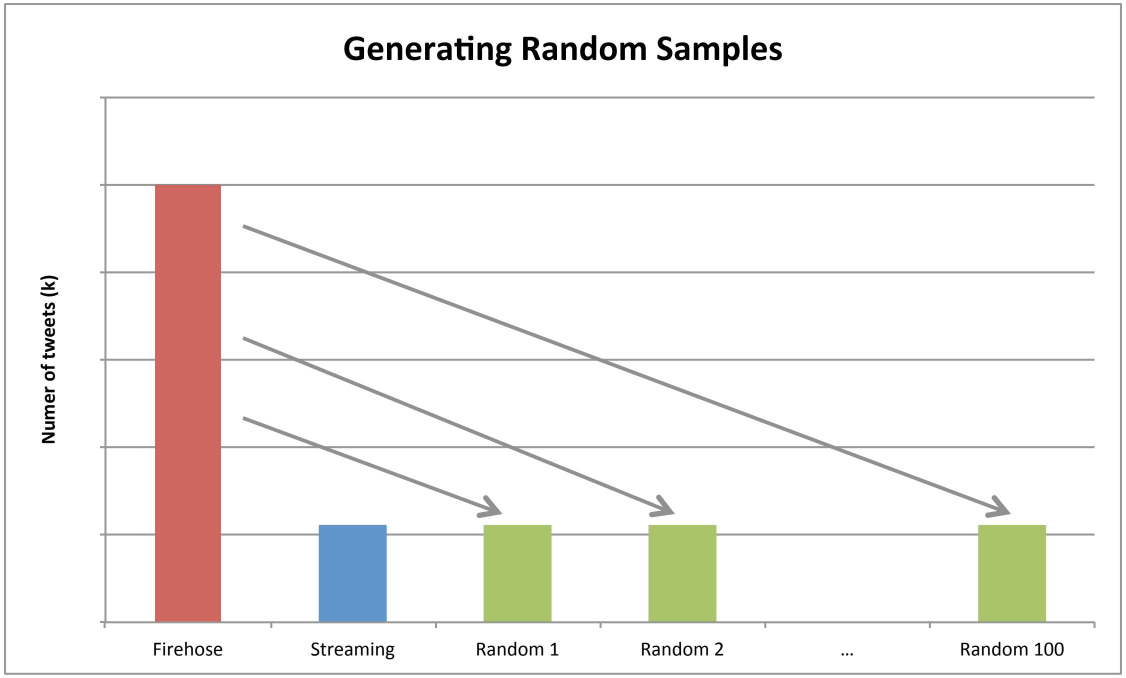


(a) Min. $\mu = 0.024$, $\sigma = 0.019$. (b) Q1. $\mu = 0.018$, $\sigma = 0.018$. (c) Median. $\mu = 0.018$, $\sigma = 0.020$. (d) Q3. $\mu = 0.014$, $\sigma = 0.016$. (e) Max. $\mu = 0.016$, $\sigma = 0.018$.

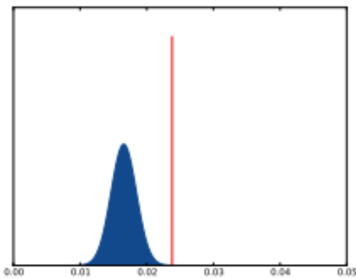
- This is just one streaming dataset against Firehose
- Are we confident about this set of results?
- Can we leverage another streaming dataset?
- Unfortunately, we cannot rewind as we have only one streaming dataset

Verification

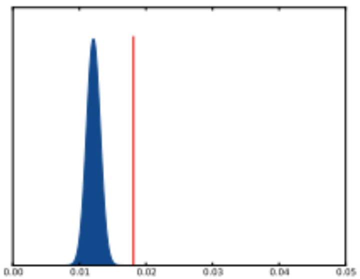
- Created 100 of our own "Streaming API" results by sampling the Firehose data.



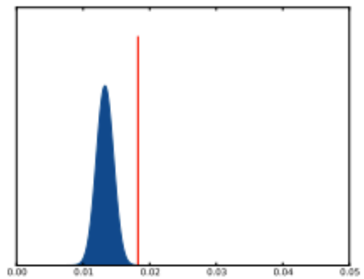
Comparison with Random Samples



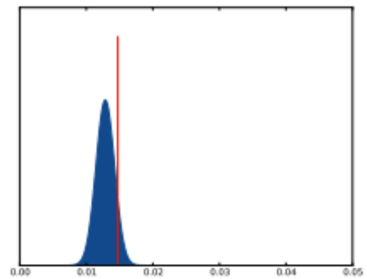
(a) Min. $S = 0.024$, $\hat{\mu} = 0.017$, $\hat{\sigma} = 0.002$, $z = 3.500$.



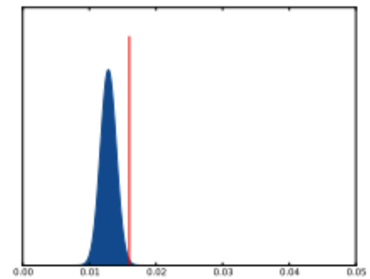
(b) Q1. $S = 0.018$, $\hat{\mu} = 0.012$, $\hat{\sigma} = 0.001$, $z = 6.000$.



(c) Median. $S = 0.018$, $\hat{\mu} = 0.013$, $\hat{\sigma} = 0.001$, $z = 5.000$.



(d) Q3. $S = 0.014$, $\hat{\mu} = 0.013$, $\hat{\sigma} = 0.001$, $z = 1.000$.



(e) Max. $S = 0.016$, $\hat{\mu} = 0.013$, $\hat{\sigma} = 0.001$, $z = 3.000$.

Summary

- Streaming API data could be biased in some facets
- Our results were obtained with the help of Firehose
- Without Firehose data, it's challenging to figure out which facets might have bias, and how to compensate them in search of credible mining results

F. Morstatter, J. Pfeffer, H. Liu, and K. Carley. *Is the Sample Good Enough? Comparing Data from Twitter's Streaming API and Data from Twitter's Firehose*. ICWSM, 2013.

Fred Morstatter, Jürgen Pfeffer, Huan Liu. *When is it Biased? Assessing the Representativeness of Twitter's Streaming API*, WWW Web Science 2014.

Some Challenges in Mining Social Media

- A Big-Data Paradox
- Studying Distrust in Social Media
- Sampling Bias
- Noise-Removal Fallacy

Noise Removal Fallacy

- We often learn that: "99% Twitter data is useless."
  - "Had eggs, sunny-side-up, this morning"
  - Can we remove noise as we usually do in DM?
- What is left after noise removal?
  - Twitter data can be rendered useless after conventional noise removal
- As we are certain there is noise in data, how can we remove it?

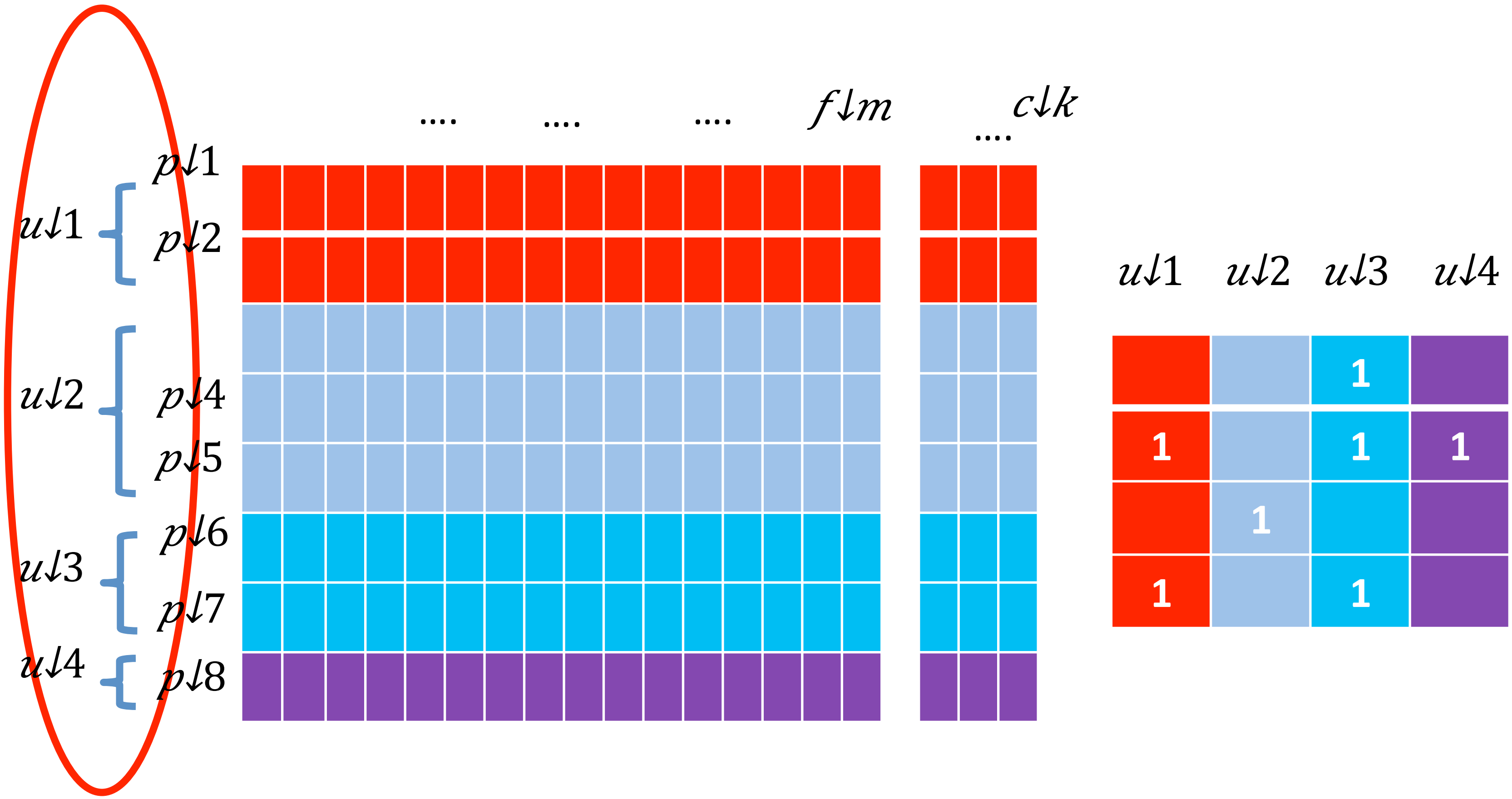
Social Media Data

- Massive and high-dimensional social media data poses unique challenges to data mining tasks
  - Scalability
  - Curse of dimensionality
- Social media data is inherently linked
  - A key difference between social media data and attribute-value data

Feature Selection of Social Data

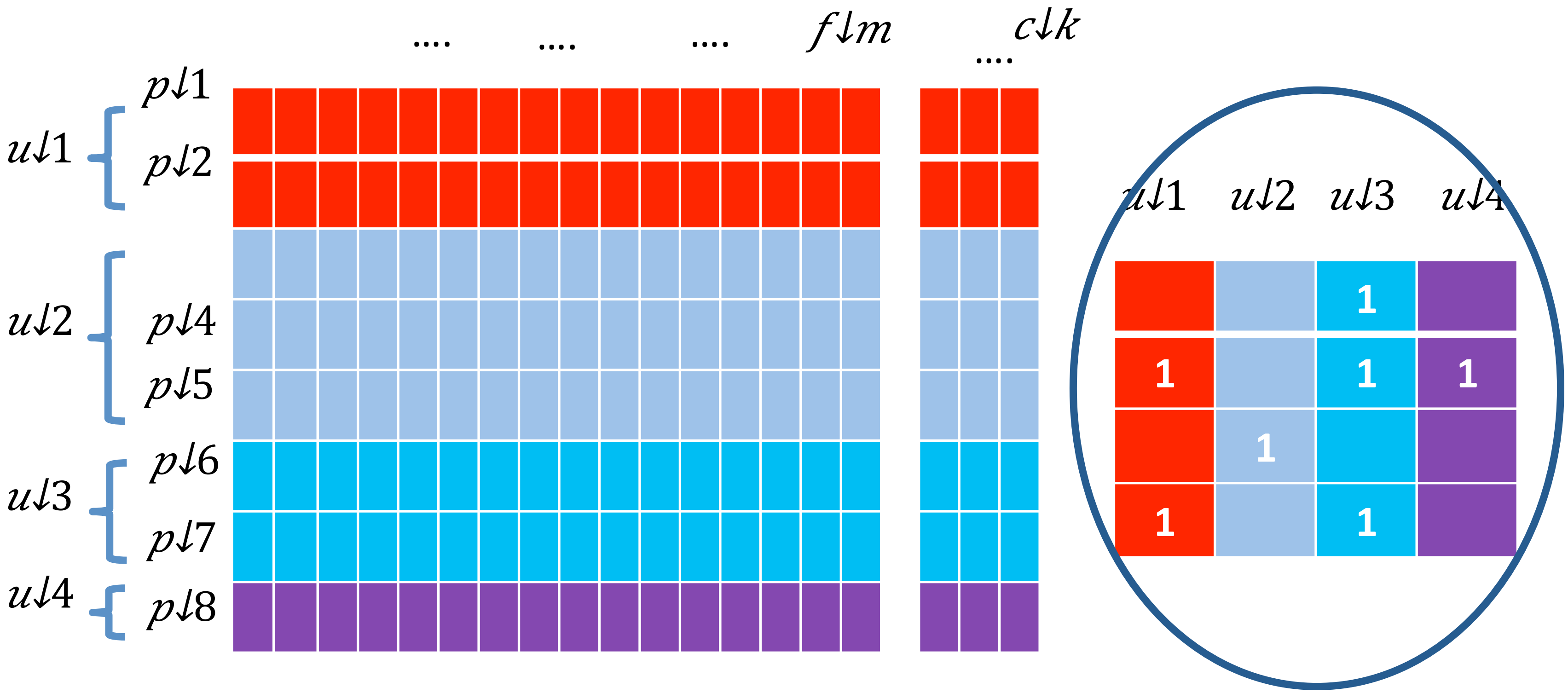
- Feature selection has been widely used to prepare large-scale, high-dimensional data for effective data mining
- Traditional feature selection algorithms deal with only “flat" data (*attribute-value data*).
  - Independent and Identically Distributed (i.i.d.)
- We need to take advantage of linked data for feature selection

Representation for Social Media Data



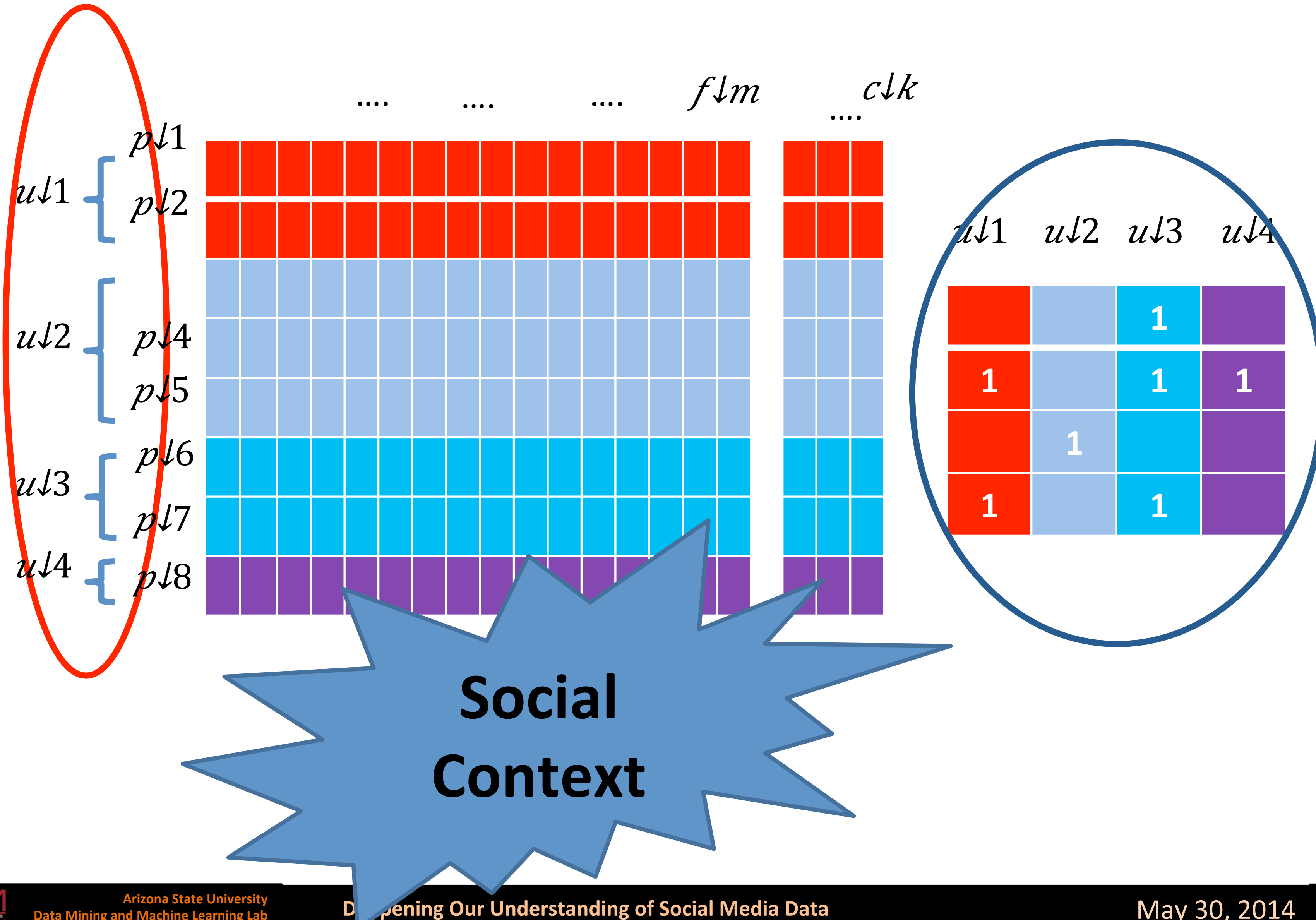
User-post relations

Representation for Social Media Data



User-user relations

Representation for Social Media Data



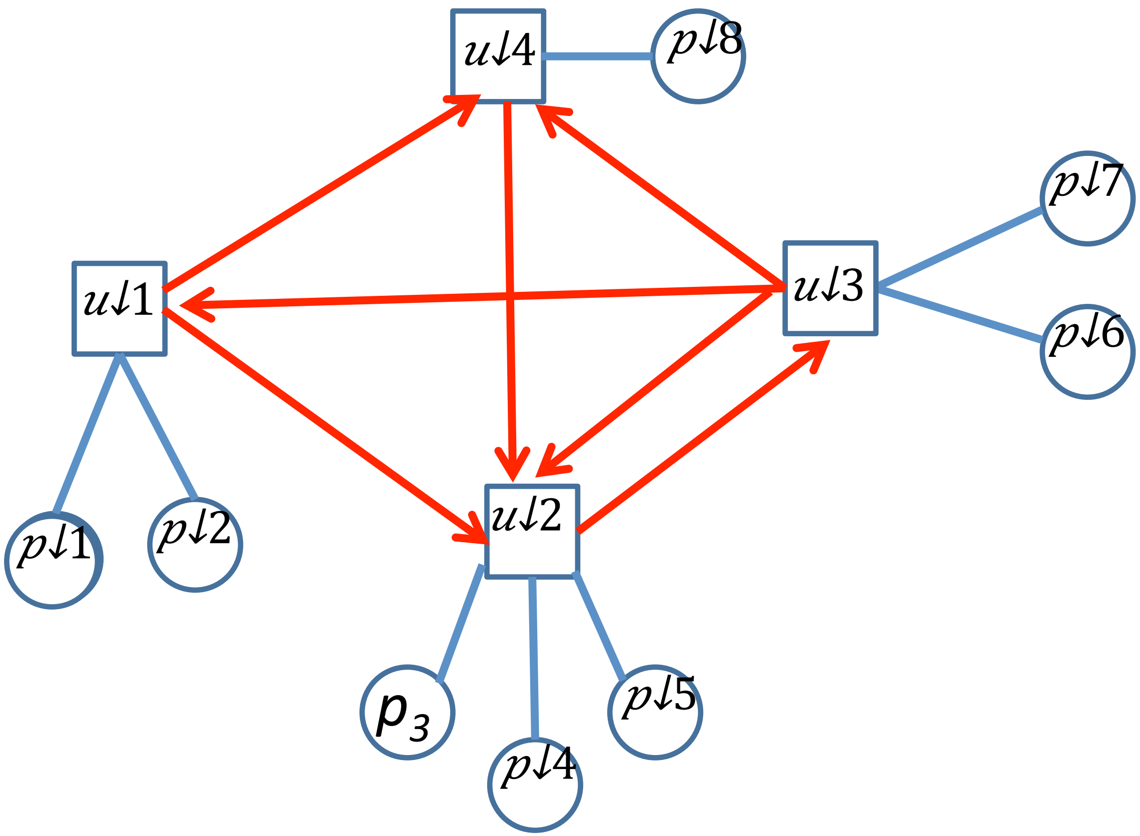
Problem Statement

- Given labeled data X and its label indicator matrix Y, the dataset F, its social context including user-user following relationships *S* and user-post relationships *P*,
- Select ***k*** most relevant features from ***m*** features on dataset *F* with its social context *S* and *P*

How to Use Link Information

- The new question is how to proceed with additional information for feature selection
- Two basic technical problems
  - Relation extraction: What are distinctive relations that can be extracted from linked data
  - Mathematical representation: How to use these relations in feature selection formulation
- Do we have theories to guide us?

Relation Extraction



1. CoPost
2. CoFollowing
3. CoFollowed
4. Following

Relations, Social Theories, Hypotheses

- Social correlation theories suggest that the four relations may affect the relationships between posts
- Social correlation theories
  - Homophily: People with similar interests are more likely to be linked
  - Influence: People who are linked are more likely to have similar interests
- Thus, four relations lead to four hypotheses for verification

Modeling CoFollowing Relation

- Two co-following users have similar topics of interests

Users' topic interests

$$\hat{T}(u_k) = \frac{\sum_{f_i \in F_k} T(f_i)}{|F_k|} = \frac{\sum_{f_i \in F_k} W^T f_i}{|F_k|}$$

$$\min_{\mathrm{W}} \| \mathrm{X}^T \mathrm{W} - \mathrm{Y} \|_F^2 + \alpha \| \mathrm{W} \|_{2,1} + \beta \sum_u \sum_{u_i, u_j \in N_u} \| \hat{T}(u_i) - \hat{T}(u_j) \|_2^2$$

Evaluation Results on Digg

Table 3: Classification Accuracy of Different Feature Selection Algorithms in Digg

Datasets	# Features	Algorithms							
		TT	IG	FS	RFS	CP	CFI	CFE	FI
$\mathcal{T}_5$	50	45.45	44.50	46.33	45.27	**58.82**	54.52	52.41	58.71
	100	48.43	52.79	52.19	50.27	**59.43**	55.64	54.11	59.38
	200	53.50	53.37	54.14	57.51	62.36	59.27	58.67	**63.32**
	300	54.04	55.24	56.54	59.27	65.30	60.40	59.93	**66.19**
$\mathcal{T}_{25}$	50	49.91	50.08	51.54	56.02	**58.90**	57.76	57.01	**58.90**
	100	53.32	52.37	54.44	62.14	64.95	64.28	62.99	**65.02**
	200	59.97	57.37	60.07	64.36	**67.33**	65.54	63.86	67.30
	300	60.49	61.73	61.84	66.80	**69.52**	65.46	65.01	67.95
$\mathcal{T}_{50}$	50	50.95	51.06	53.88	58.08	59.24	59.39	56.94	**60.77**
	100	53.60	53.69	59.47	60.38	65.57	64.59	61.87	**65.74**
	200	59.59	57.78	63.60	66.42	70.58	68.96	67.99	**71.32**
	300	61.47	62.35	64.77	69.58	77.86	71.40	70.50	**78.65**
$\mathcal{T}_{100}$	50	51.74	56.06	55.94	58.08	**61.51**	60.77	59.62	60.97
	100	55.31	58.69	62.40	60.75	63.17	63.60	62.78	**65.65**
	200	60.49	62.78	65.18	66.87	**69.75**	67.40	67.00	67.31
	300	62.97	66.35	67.12	69.27	**73.01**	70.99	69.50	72.64

Evaluation Results on Digg

Table 3: Classification Accuracy of Different Feature Selection Algorithms in Digg

Datasets	# Features	Algorithms							
		TT	IG	FS	RFS	CP	CFI	CFE	FI
$\mathcal{T}_5$	50	45.45	44.50	46.33	45.27	**58.82**	54.52	52.41	58.71
	100	48.43	52.79	52.19	50.27	**59.43**	55.64	54.11	59.38
	200	53.50	53.37	54.14	57.51	62.36	59.27	58.67	**63.32**
	300	54.04	55.24	56.54	59.27	65.30	60.40	59.93	**66.19**
$\mathcal{T}_{25}$	50	49.91	50.08	51.54	56.02	**58.90**	57.76	57.01	**58.90**
	100	53.32	52.37	54.44	62.14	64.95	64.28	62.99	**65.02**
	200	59.97	57.37	60.07	64.36	**67.33**	65.54	63.86	67.30
	300	60.49	61.73	61.84	66.80	**69.52**	65.46	65.01	67.95
$\mathcal{T}_{50}$	50	50.95	51.06	53.88	58.08	59.24	59.39	56.94	**60.77**
	100	53.60	53.69	59.47	60.38	65.57	64.59	61.87	**65.74**
	200	59.59	57.78	63.60	66.42	70.58	68.96	67.99	**71.32**
	300	61.47	62.35	64.77	69.58	77.86	71.40	70.50	**78.65**
$\mathcal{T}_{100}$	50	51.74	56.06	55.94	58.08	**61.51**	60.77	59.62	60.97
	100	55.31	58.69	62.40	60.75	63.17	63.60	62.78	**65.65**
	200	60.49	62.78	65.18	66.87	**69.75**	67.40	67.00	67.31
	300	62.97	66.35	67.12	69.27	**73.01**	70.99	69.50	72.64

Summary

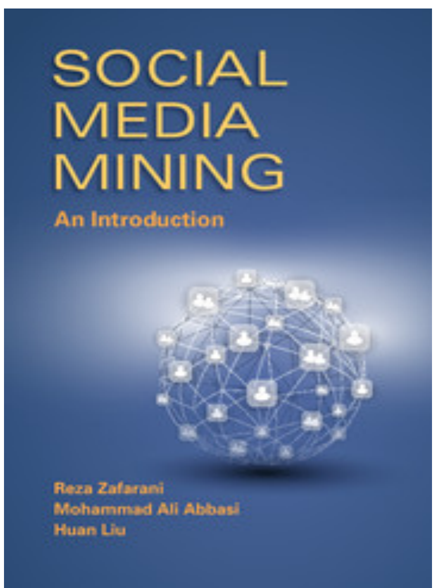
- LinkedFS is evaluated under varied circumstances to understand how it works.
  - Link information can help *feature selection for social media data.*
- Unlabeled data is more often in social media, unsupervised learning is more sensible, but also more challenging.

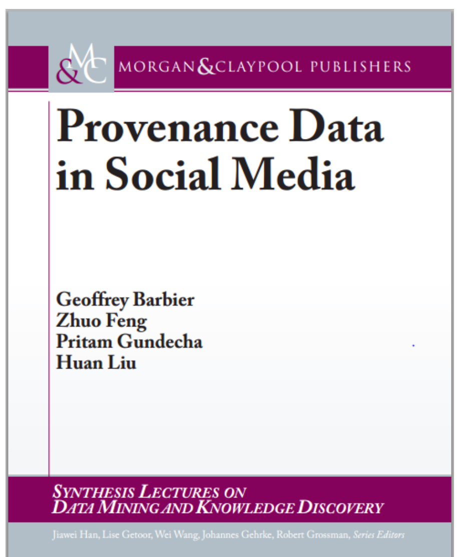
Jiliang Tang and Huan Liu. `` Unsupervised Feature Selection for Linked Social Media Data'', the Eighteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining , 2012.

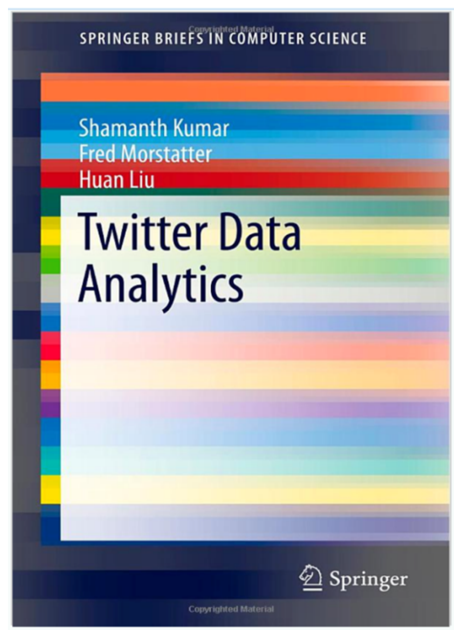
Jiliang Tang, Huan Liu. ``Feature Selection with Linked Data in Social Media'', SIAM International Conference on Data Mining, 2012.

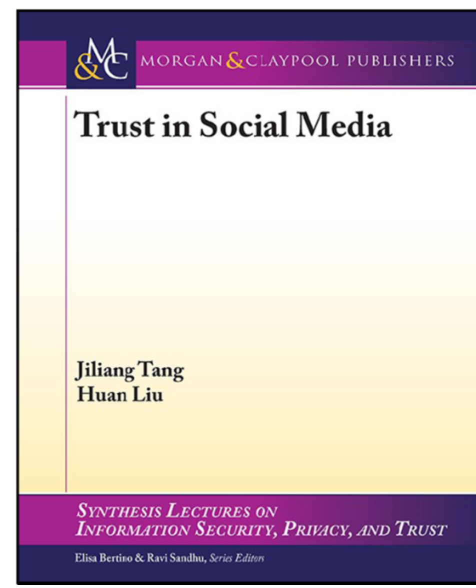
Concluding Remarks

- A Big-Data Paradox
- Studying Distrust in Social Media
- Sampling Bias in Social Media Data
- Noise Removal Fallacy









THANK YOU and THANKS to …

- Organizers for this wonderful opportunity to share our research work
- Acknowledgments
  - Grants from NSF, ONR, ARO
  - DMML members and project leaders
  - Collaborators


Shamanth Kumar


Ali Abbasi


Reza Zafarani


Fred Morstatter

Jiliang Tang