The Good, the Bad and the Ugly
- Uncovering Novel Opportunities of Data Science

Huan Liu

Social Media Mining

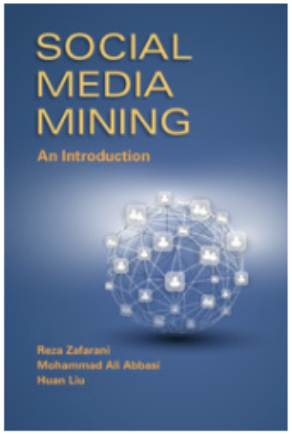
Home | Download Book | Slides/Tutorials | Table of Contents | Errata | How to Cite

Social Media Mining

An Introduction

A Textbook by Cambridge University Press

Reza Zafarani	*Syracuse University*
Mohammad Ali Abbasi	*Machine Zone*
Huan Liu	*Arizona State University*



PDF DOWNLOAD

Accessed 90,000+ times
from 160+ countries and 1200+ Universities

CAMBRIDGE UNIVERSITY PRESS | amazon.com | BARNES & NOBLE BOOKSELLERS | eBooks.com | TURING

The growth of social media over the last decade has revolutionized the way individuals interact and

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

Big Data Challenges Traditional Thinking

- Data is ubiquitous and can only become bigger
- Big data is not just big
  - Transforming how we live, work, and think
- Big data makes many tasks easier and better
- An example of big mobile data
  - Using GPS to guide our travel *today* vs. *not so long ago*
- Opportunities are where challenges are

Traditional Media and Data

Broadcast Media
One-to-Many

Communication Media
One-to-One

Traditional Data

Some Challenges in Understanding Social Media

- Noise-Removal Fallacy
  - Can we remove noise without losing much information?
- Studying Distrust (the Implicit) in Social Media
  - Where to find the invisible distrust?
- Big-Data Paradox
  - Lack of data with big social media data
- Evaluation Dilemma
  - Where is ground truth? How to evaluate without it?
- Data Sampling Bias and Its Mitigation
  - Often we get a small sample of (still big) data. Would that data suffice to obtain credible findings?

The Good, the Bad, and the Ugly of Social Media Data

- The **good**
  - Social media data is big and linked
- The **bad**
  - Social media data is noisy and short of data where it is most needed
- The **ugly**
  - Social media data is heterogeneous, partial, and asymmetrical

Two Illustrative Cases for *Novel Challenges*:

(1) Removing noise, and
(2) Inferring the implicit

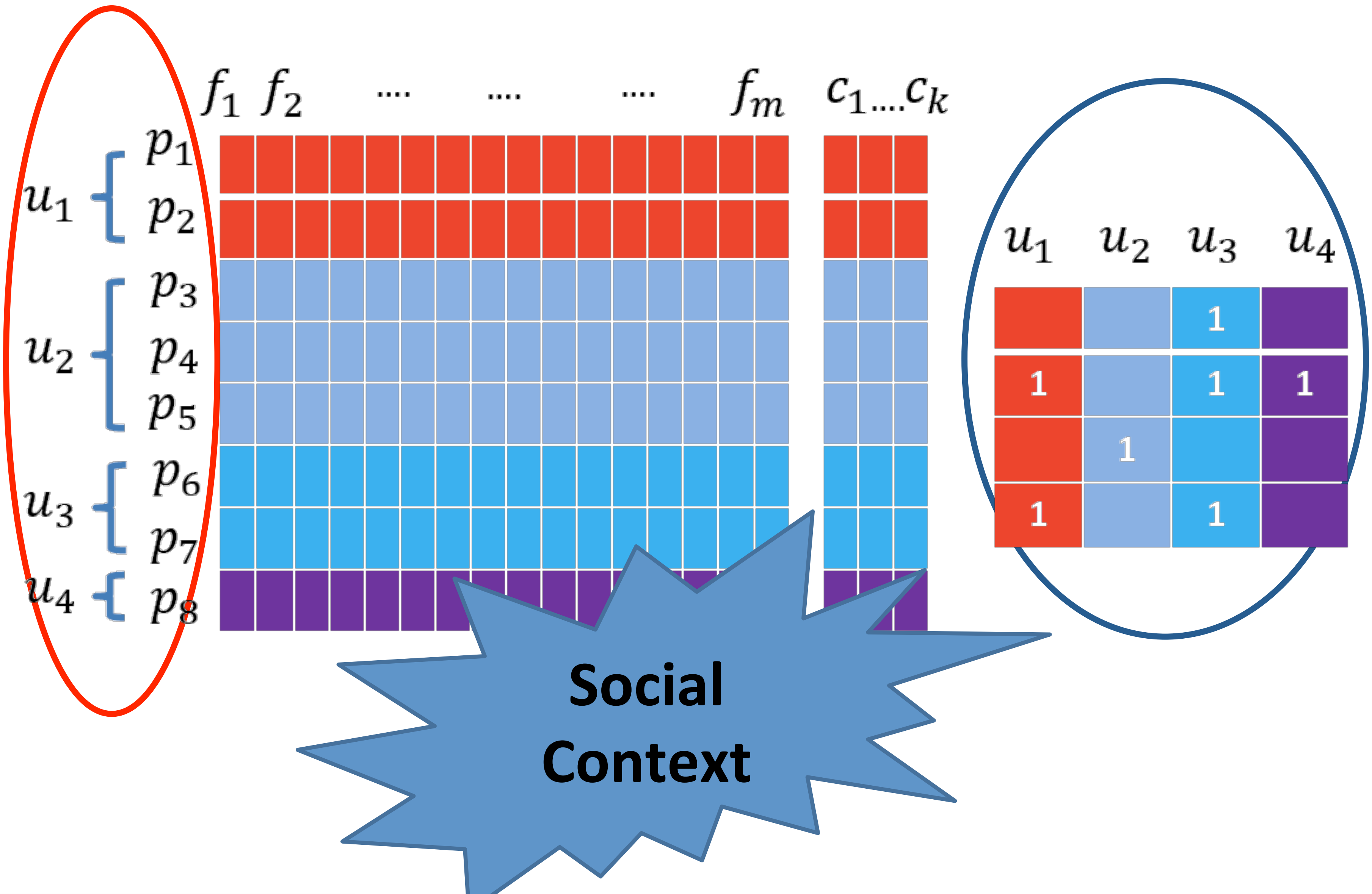
Removing Noise – a First Task in Data Mining

- We often heard 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, should we remove it?
  - If *yes*, how?
- **A new challenge**: Feature selection with linked data

Social Data and Feature Selection

- High-dimensional social media data poses unique challenges to data mining tasks
- 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*).
- We now can take advantage of *linked* data for feature selection

Representation for Social Media Data



New Problem Statement of Feature Selection

- 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 for dataset *F* with its social context *S* and *P*

How to Use Link Information

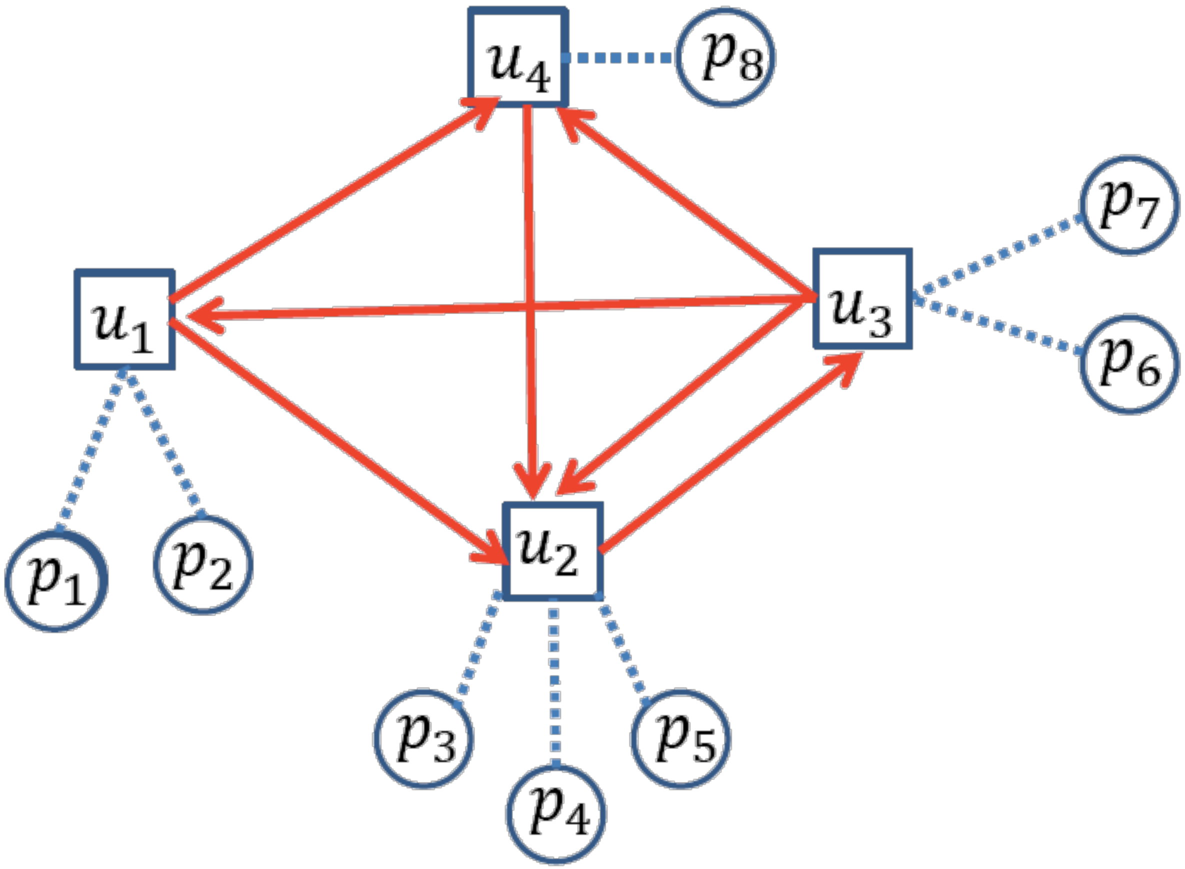
- Would the additional (i.e., link) information be useful for feature selection?
- Some technical challenges
  - Relation extraction: What are distinct relations that can be extracted from linked data
  - Mathematical representation: How to use these relations in feature selection formulation
- Are there theories to guide us in generating hypotheses?

Social Theories Guided Research

- 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
- Guided by theories, we turn

social relations → hypotheses for investigation

Relation Extraction



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

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

- We evaluate if link information can be used for feature selection and understand how it works
  - Link information can help *feature selection for social media data,* in particular, when we are *short of* data
- *Unlabeled* data is more often in social media, unsupervised learning is more sensible, but also more challenging

Inferring the Implicit – Second Case

- Both trust and distrust (positive and negative info) help decision makers reduce the uncertainty and risk associated with decisions
- Distrust may play an equally, if not more, critical role as trust does in decision making

- Distrust is *new* in Social Media Analysis
  - Asymmetry of information available (like vs dislike)
- Distrust is, however, *not new* in Social Sciences
  - Various definition of distrust in Social Sciences

Two Theories 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

Challenges in Studying Distrust 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 that social sciences conventionally use to conduct studies
- Challenge 2: Distrust information is usually not publicly available
  - Trust is desired while distrust is not for open online social platforms

Computational Understanding of Distrust

- Design computational tasks to help understand distrust with *passively observed* social media data
  - Q1: **Is distrust the negation of trust?**
    - Yes or No?
  - Q2: **Is there any value of distrust after Q1 is answered?**
    - If distrust is a new dimension of trust, what is added value of distrust
- How can we use social media data to computationally answer the two questions?

Task 1: Is distrust the negation of trust?

- If distrust is the negation of trust, or low trust is equivalent to distrust, distrust should be predictable using trust information

IF Distrust ≡ Low Trust

THEN Predicting Distrust ≡ Predicting Low Trust

Evaluation of Task 1

- The performance of using low trust for distrust is consistently worse than randomly guessing
- Task 1: Since it fails to predict distrust with only trust, 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: Is there any added value of distrust?

- If distrust has any added value, we should predict trust better with distrust

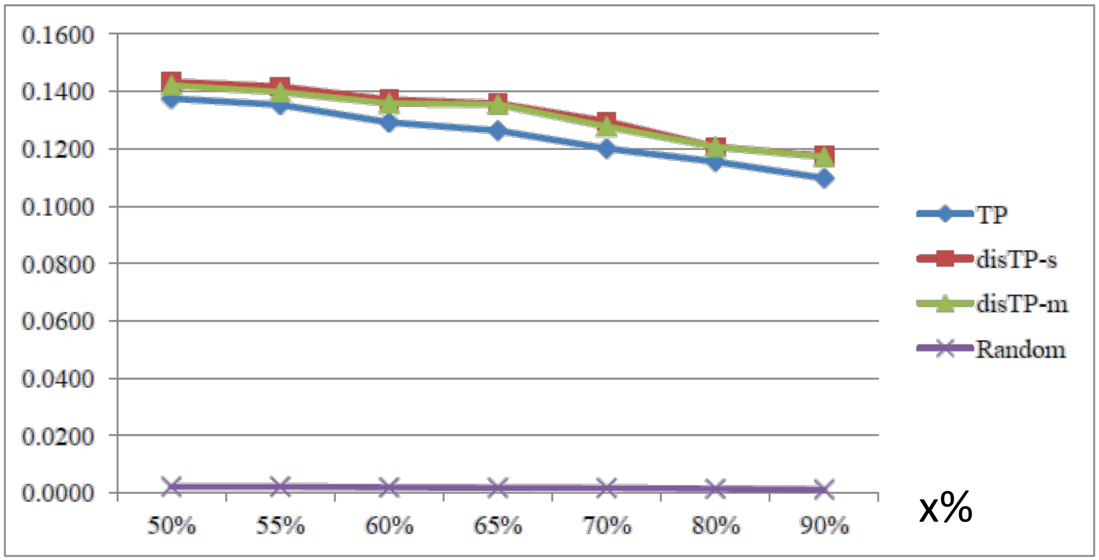
- To verify the above statement, we define the second computational task involving distrust
  - Incorporating distrust in **trust prediction**

Evaluation of Distrust in Trust Propagation

- Incorporating distrust propagation can improve the performance of trust measurement

	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

PA Performance

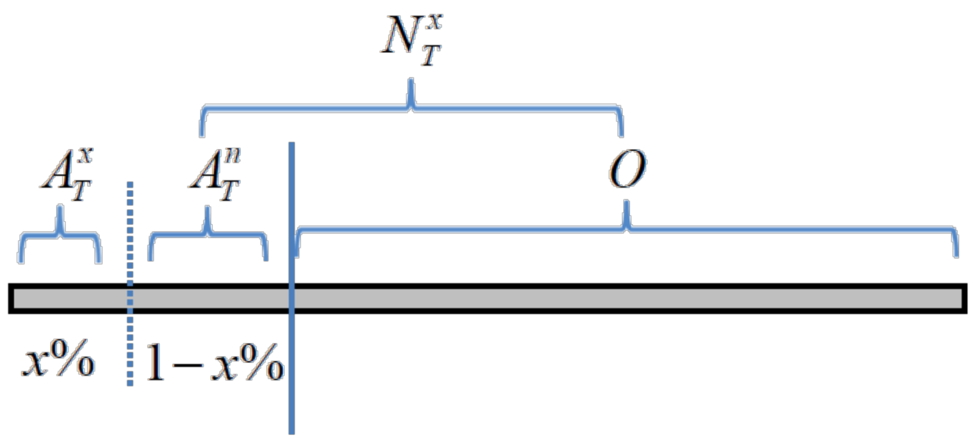


x

- One step distrust propagation usually outperforms multiple step distrust propagation

Experimental Settings for Task 2

- x% of pairs of users with trust relations are chosen as old trust relations and the remaining as new trust relations



- Task 2 predicts $|A_T^n|$ pairs of users P from N_T^x as new trust relations

- The performance is computed as $PA = \frac{|A_T^n \cap P|}{|A_T^n|}$

Findings from Understanding Distrust

- Distrust presents distinct properties
  - Properties of trust cannot be extended to distrust
- Distrust is not the negation of trust
  - Low trust fails to predict distrust
- Distrust has added value over trust
  - Distrust helps improve trust prediction performance

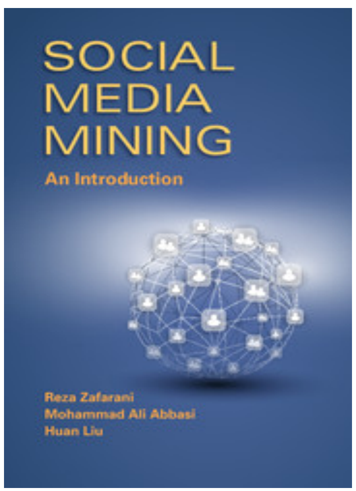
- However, distrust information is usually not available on a social networking site
- Next task - discovering negative links like distrust

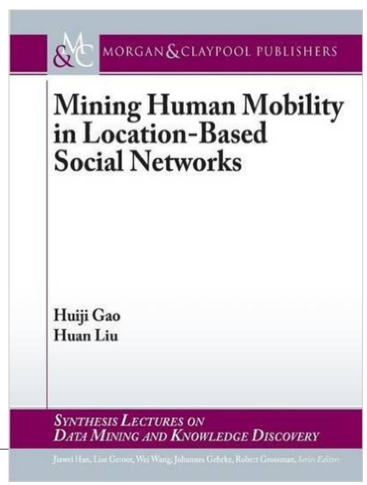
Some Challenges in Understanding Social Media

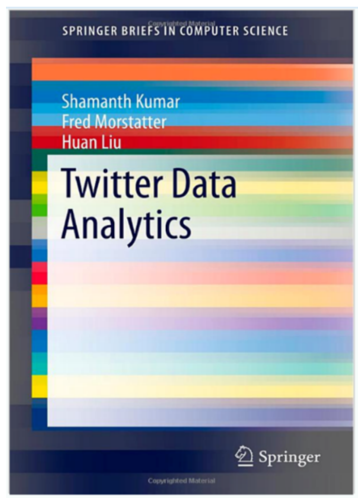
- Noise-Removal Fallacy
  - Can we remove noise without losing much information?
- Studying Distrust in Social Media
  - Where to find the invisible distrust?
- Big-Data Paradox
  - Lack of data with big social media data
- Evaluation Dilemma
  - Where is ground truth? How to evaluate without it?
- Sampling Bias and Its Mitigation
  - Often we get a small sample of (still big) data. Would that data suffice to obtain credible findings?

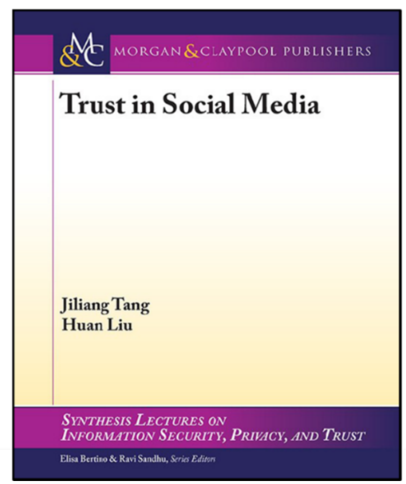
Repositories and Recent Books

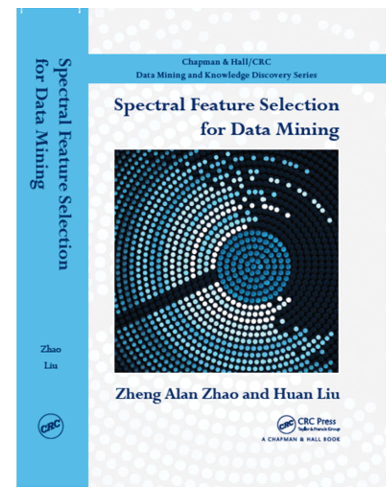
- ***scikit-feature*** – an open source feature selection repository in Python
- Social Computing Repository
- Some books available as free download











THANK YOU and DFC2016

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

Search "huan Liu" for more information or at http://www.public.asu.edu/~huanliu

H Liu, F Morstatter, J Tang, and R Zafarani. ``**The good, the bad, and the ugly: uncovering novel research opportunities in social media mining",** in Trends of Data Science, International Journal on Data Science and Analytics, Springer International Publishing Switzerland. September, 2016. DOI 10.1007/s41060-016-0023-0