# Mining Social Media: Looking Ahead

Huan Liu

Joint work with

DMML Members and Collaborators

http://dmml.asu.edu/

2014.10.22: Dr. H. Russell Bernard and Dr. Lisa Troyer Visit DMML Group@ASU

# Social Media Mining by Cambridge University Press

## Social Media Mining

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

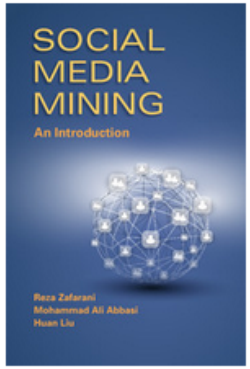
# Social Media Mining

## *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/**

# Social Media Data Is Big

- Social media is a key player in the Big Data era
  - We're overwhelmed by the data, start appreciating data value, and data is ubiquitous
  - Social media data is a new species
- Big data will only become bigger
  - Rapid growth of linked data (MOOCS, eCommerce, IOT)
- Big data is a good problem to have
  - And, many a time, big data may not be big enough

# Big Social Media Data, Big and New Challenges

- A Big-Data Paradox
  - Still lack of data with big social media data
- Trust vs. Distrust in Social Media
  - Is distrust information important? Where can we find distrust information with "one-way" relations?
- Data Sample Sufficiency
  - Often we get a small sample of (still big) data. Would that sample suffice to obtain credible findings?
- Evaluation Dilemma
  - Where is ground truth? How to evaluate w/o it?

# A Big-Data Paradox

- Collectively, social media data is indeed big
- Individually, however, the data is *little*
  - How much activity data do we generate daily?
  - How many posts did we post this week?
  - How many friends do we have?
- Our data also appears at various social media sites as we use them for different purposes
  - WeChat, Facebook, Twitter, Instagram, …
- When "big" social media data isn't big enough,
  - Searching for **more** data with **little** data

# An Example

- **Little** data about an individual

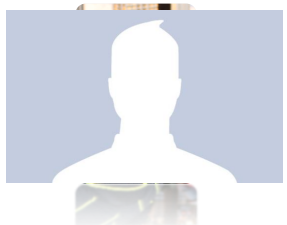
+ **Many** social media sites

- **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**

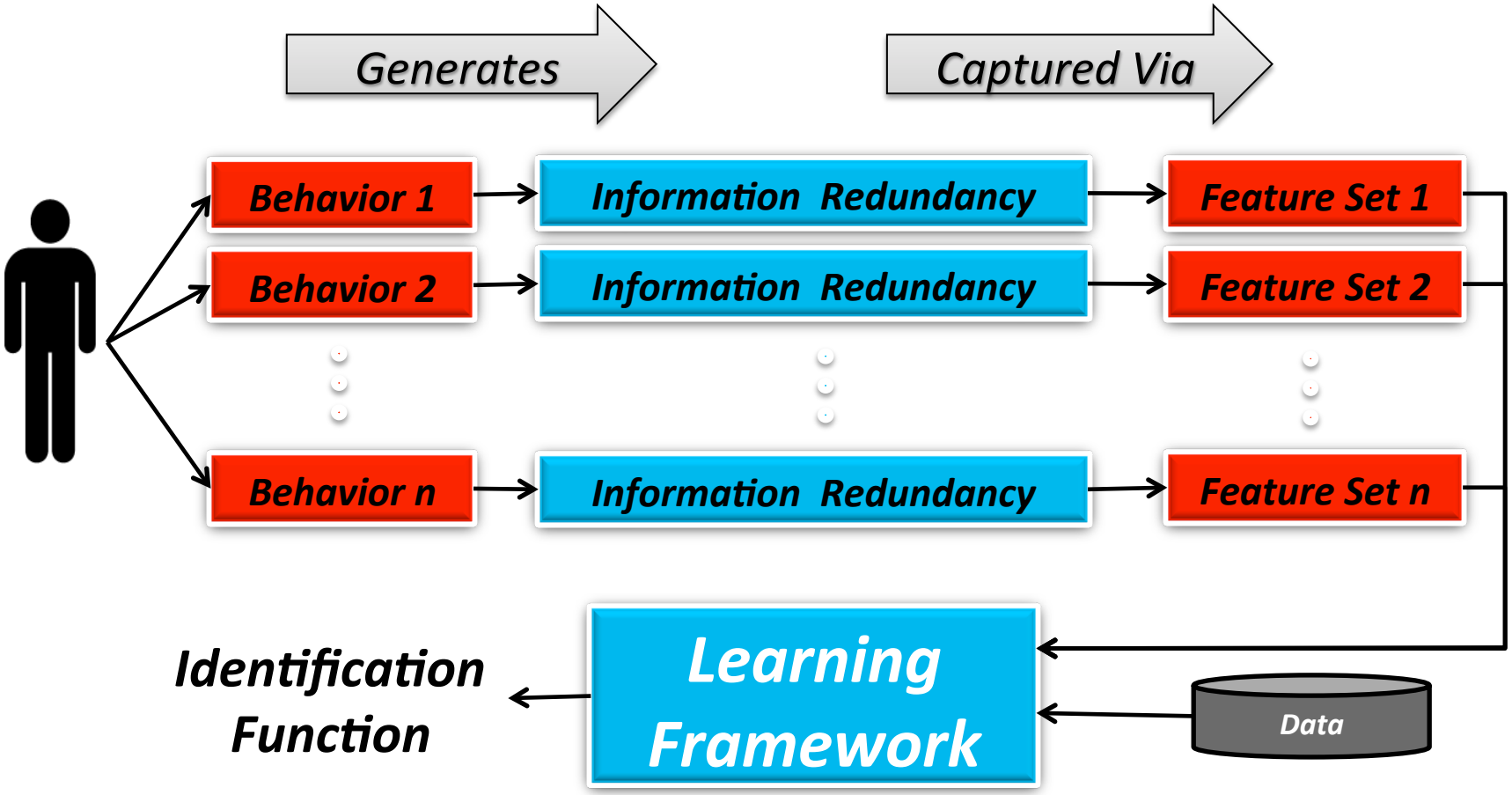
***Can we connect individuals across sites?***

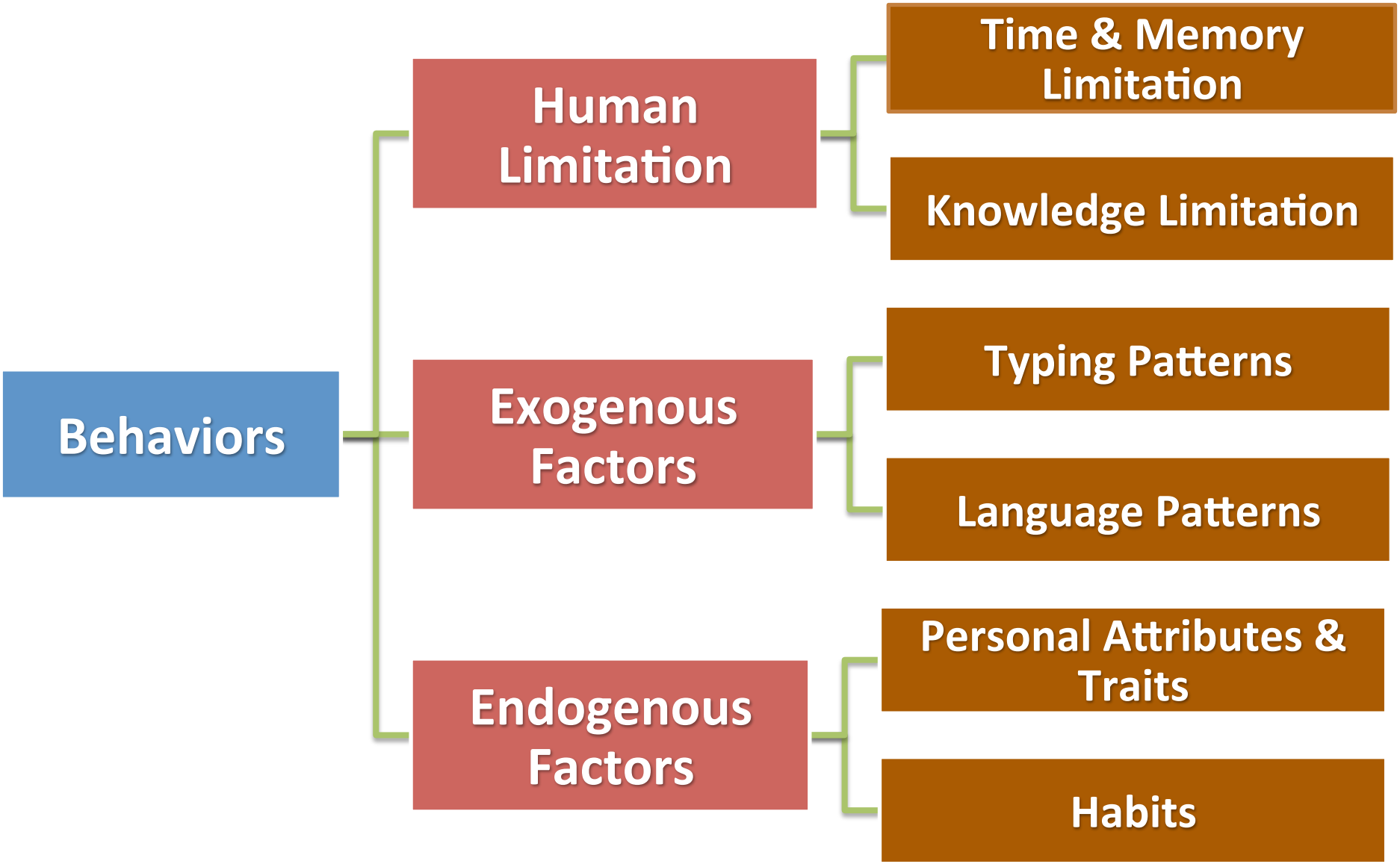
Reza Zafarani and Huan Liu. ``Connecting Users across Social Media Sites: A Behavioral-Modeling Approach", the Nineteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'2013), August 11 - 14, 2013. Chicago, Illinois.

# Searching for More Data with Little Data

- Each social media site can have varied amount of user information
- Which information definitely exists for all sites?
  - **Usernames**
  - But, a user's usernames on different sites can be different
- Our work is to verify if the information provided across sites belong to the same individual

# A Behavioral Modeling Approach with Learning




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

# 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

# Summary

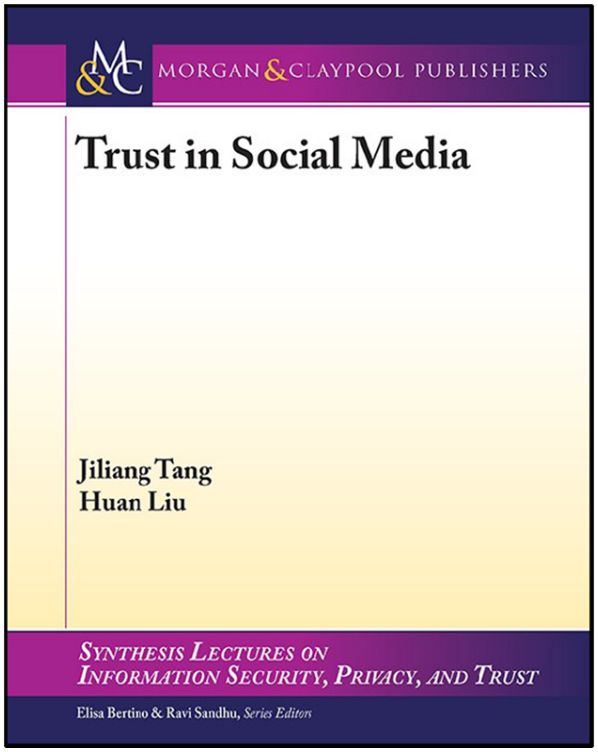
- Many a time, big data may not be sufficiently big for a data mining task
- Gathering more data is often necessary for effective data mining
- Social media data provides unique opportunities to do so by using numerous sites and abundant user-generated content
- Traditionally available data can also be tapped to make “thin” data “thicker”

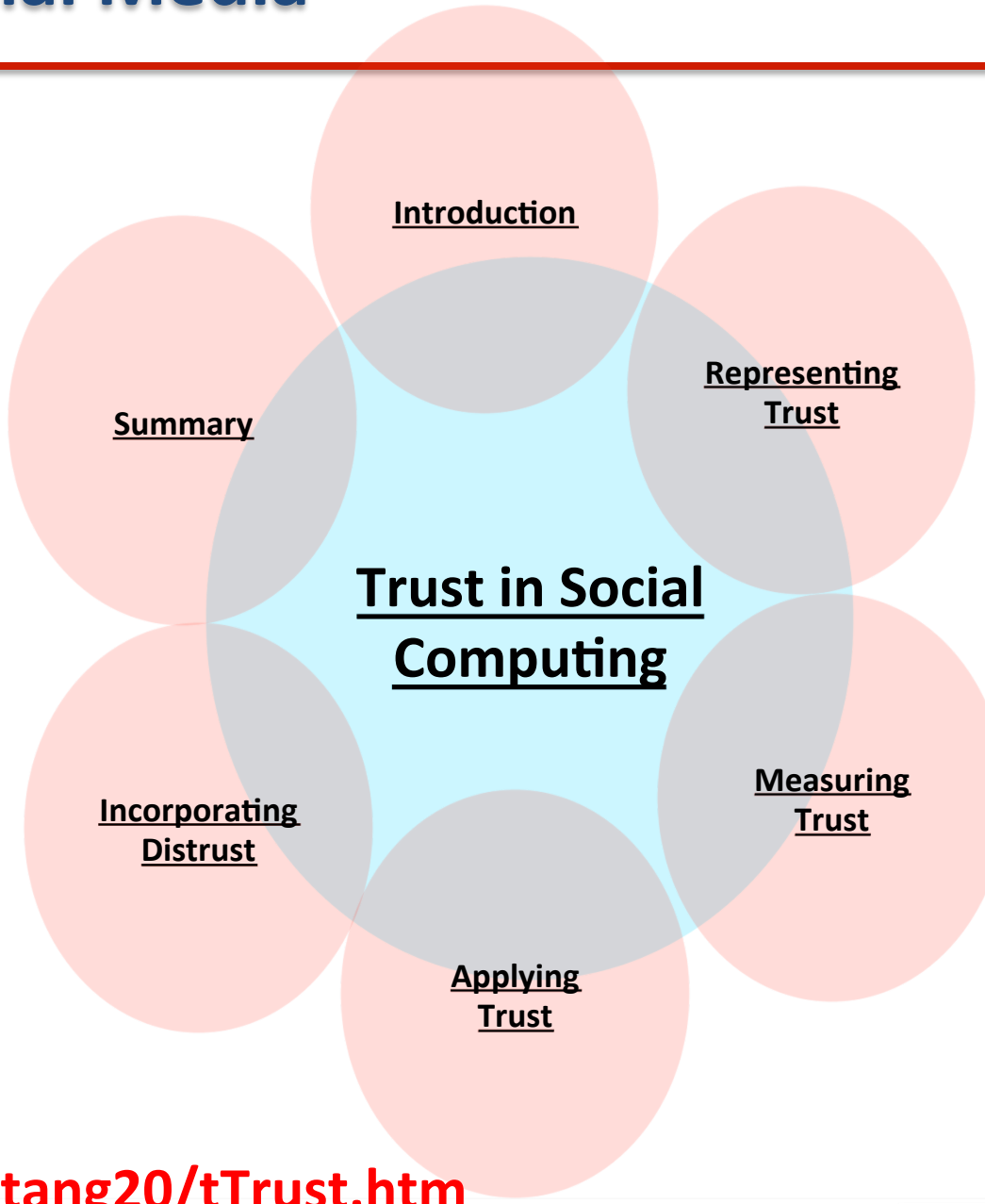
Reza Zafarani and Huan Liu. ``Connecting Users across Social Media Sites: A Behavioral-Modeling Approach", SIGKDD, 2013.

# New Challenges in Mining Social Media

- A Big-Data Paradox
- Trust vs. Distrust in Social Media
- Data Sample Sufficiency
- Evaluation Dilemma

# 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

# Understanding 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

Jiliang Tang, Xia Hu, and Huan Liu. ``Is Distrust the Negation of Trust? The Value of Distrust in Social Media", 25th ACM Conference on Hypertext and Social Media (HT2014), Sept. 1-4, 2014, Santiago, Chile.

# 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, we can just use trust information alone
- Task 2: **Is distrust information valuable?**
  - If distrust is a new dimension of trust, does distrust have added value
- The first step to understand distrust is to make distrust computable in trust models

# A Computational Understanding of Distrust

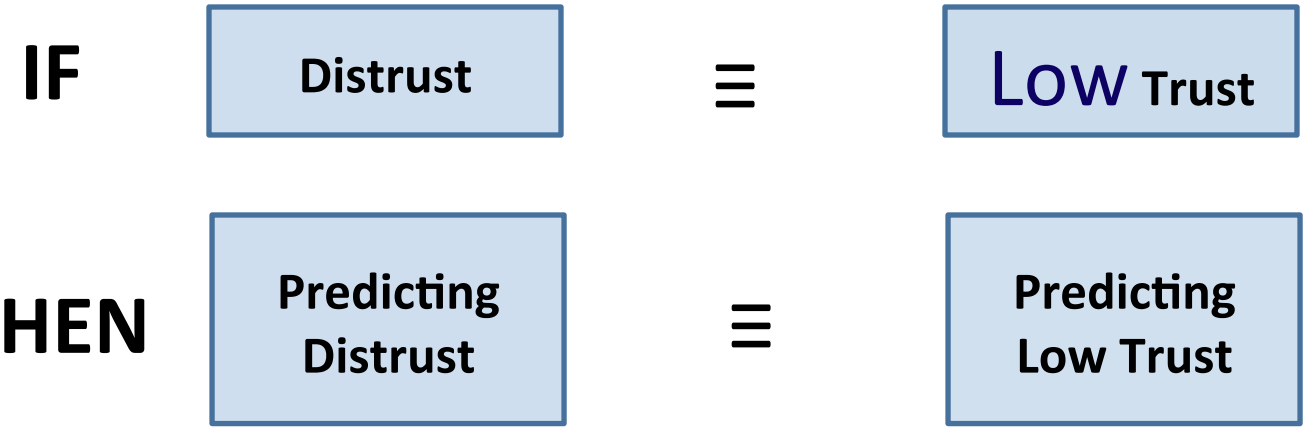
- Social media data is a new type of social data
  - Passively observed
  - Large scale

- Task 1: Is distrust the negation of trust?
  - Predicting distrust from only trust

- Task 2: Does distrust have added value on trust?
  - Predicting trust with distrust

# 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




- 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 may provide additional information about users, and could have added value beyond trust

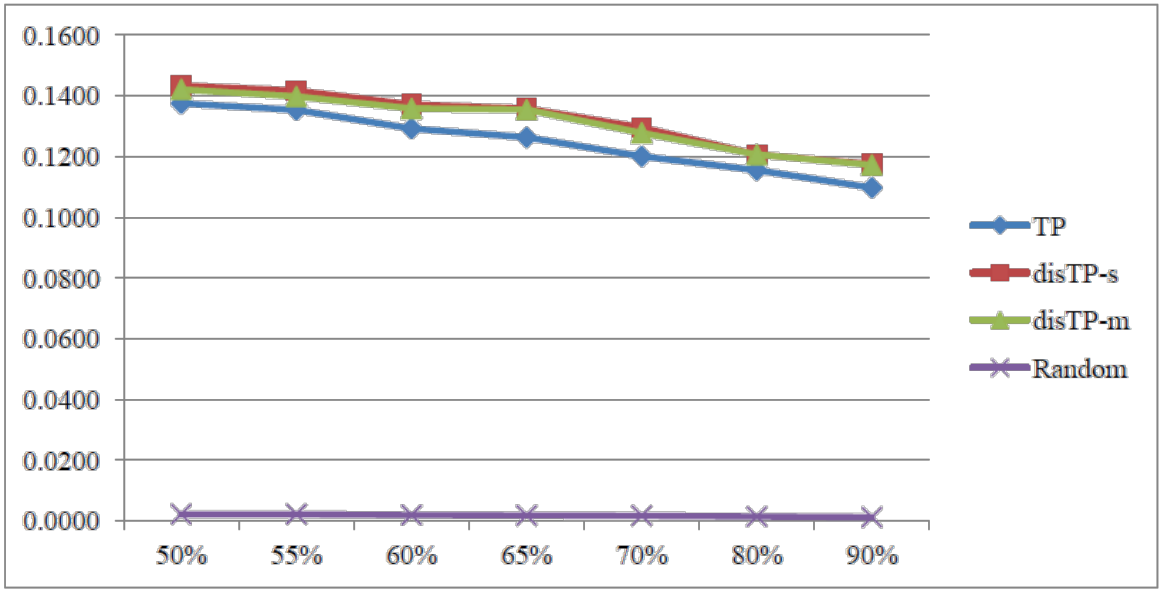
- We seek answer to the questions - whether using both trust and distrust information can help achieve better performance than using only trust information

- 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 |



# Findings from the Computational Understanding

- Task 1 shows that distrust is not the negation of trust
  - Low trust is not equivalent to distrust
- Task 2 shows that trust can be better measured by incorporating distrust
  - Distrust has added value in addition to trust
- This computational understanding suggests that it is necessary to compute distrust in social media
- What are the next steps of distrust research?

# New Challenges in Mining Social Media

- A Big-Data Paradox
- Trust vs. Distrust in Social Media
- Data Sample Sufficiency
- Evaluation Dilemma

# THANK YOU …

- For this wonderful opportunity for sharing
- Acknowledgments
  - Grants from NSF, ONR, and ARO
  - DMML members and project leaders
  - Interdisciplinary collaborators

- **Summary**
  - A Big-Data Paradox
  - Trust vs. Distrust
  - *Data Sample Sufficiency*
  - *Evaluation Dilemma*

# Concluding Remarks

- Social media offers great opportunities
- There are many challenging problems
- Exciting time for science, engineering, & business