• **Huan Liu** is a professor of computer science and engineering at Arizona State University. His research interests are in data and web mining, machine learning, social computing, artificial intelligence, and investigating problems that arise in many real-world applications with high-dimensional data of disparate forms. He is a member of the AAAI, ACM, and the American Society for Engineering Education, and is a fellow of IEEE. For contact information and links to recent publications, please visit [http://www.public.asu.edu/~huanliu/](http://www.public.asu.edu/~huanliu/).

• **Pritam Gundecha** is a computer science Ph.D. student at Arizona State University. His research interests include data mining, machine learning and social computing. His research focuses on security, privacy and trust issues in social media. For contact information and links to recent publications, please visit [http://www.public.asu.edu/~pgundech/](http://www.public.asu.edu/~pgundech/).
Mining Social Media: A Brief Introduction

Pritam Gundecha and Huan Liu

A Tutorial in Operations Research
INFORMS 2012
Social Media is the use of electronic and Internet tools for the purpose of sharing and discussing information and experiences with other human beings in more efficient ways.
Social Media Landscape
Social Media Examples

• A wiki article
• Web reviews and ratings of a popular hamburger restaurant in your city
  – E.g., Yelp.com
• An online social network of your professional contacts
  – E.g., Facebook.com, LinkedIn.com
• A smart phone app that informs you about the police and camera locations in a street
  – “The Web is dead.” – Chris Anderson of wired.com
Types of Social Media

- Online Social Networking
  - Publishing
    - Blogging, Wiki, etc.
- Micro blogging
- Social News
- Social Bookmarking
- Media Sharing
  - Video, Photo, and Podcast Sharing
- Opinion, Review, and Ratings Websites
- Answers
- Entertainment
Online Social Networks are web-based services that allow individuals and communities to connect with real world friends and acquaintances online.

- **Interactions**
  - Friendship interaction
  - Friends, like, comments, ...
  - Media Sharing
  - Sending and receiving messages

- **Examples**
  - Facebook.com
  - MySpace.com
  - Bebo.com
  - Orkut.com
Blogging

A blog is a journal-like website for users, a.k.a. bloggers, to contribute textual and multimedia content, arranged in reverse chronological order

- Maintained both individually or by a community (See a tutorial at KDD http://videolectures.net/kdd08_liu_briat/)
- Usages:
  - Sharing information and opinions with friends and strangers
  - Disseminating subject-specific content
  - Who is the influential http://videolectures.net/wsdm08_agarwal_iib/
Microblogging is a relatively new phenomenon that can be considered as a counterpart to blogging, but with limited content.

- **Usage**
  - communication medium
  - social interaction
  - citizen journalism

- **Service Providers:**
  - Twitter
  - Weibo
A wiki is a collaborative editing environment that allows users to develop Web pages using a simplified markup language

• Wikipedia allows interested individuals to collaboratively develop articles on a variety of subjects.
• Using the wisdom of crowds effectively, it has become a comprehensive repository of information useful to a variety of individuals.
Social News refers to the sharing and selection of news stories and articles by a community of users.

- Users can share articles that they believe would interest the community
- Samples:
  - Digg.com
  - Slashdot
  - Fark
  - Reddit
Social Bookmarking sites allow users to bookmark web content for storage, organization and sharing.

- These bookmarks can be tagged with metadata to categorize and provide context to the shared content, allowing users to organize information making it easy to search and identify relevant information.

- Samples
  - Delicious.com
  - StumbleUpon.com
Media sharing

Media sharing is an umbrella term that refers to the sharing of a variety of media on the web. Users share such multimedia content of possible interest to others

• Samples:
  – Video Sharing: YouTube.com
  – Photo Sharing: Flicker.com, picasa.com
  – Document Sharing: Scribd.com, Slidesharec.com
  – Livecasting: Justin.tv, Ustream.com
Opinion, review, and ratings websites are websites whose primary function is to collect and publish user-submitted content in the form of subjective commentary on existing products, services, entertainment, businesses, places, etc. Some commercial sites may serve a secondary purpose as review sites by publishing product reviews submitted by customers.

- Examples
  - Cnet.com
  - Epinions.com
  - yelp.com
  - tripadvisor.com
Socially Improved Answers

In these sites, users who require certain guidance, advice or knowledge can ask questions. Other users from the community can answer these questions based on knowledge acquired from previous experiences, personal opinions or from relevant research.

– Unlike review and opinion sites, which contain self-motivated contribution of opinions, answer sites contain knowledge shared in response to a specific query.

• Samples:
  – WikiAnswers, Yahoo Answers
Main Characteristics

• Participation
  – Social media encourages contributions and feedback from everyone who is interested. It blurs the line between media and audience.

• Openness
  – Most social media services are open to feedback and participation. They encourage voting, comments and the sharing of information. There are rarely any barriers to accessing and making use of content – password-protected content is frowned on.

• Conversation
  – Whereas traditional media is about “broadcast” (content transmitted or distributed to an audience) social media is better seen as a two-way conversation.

• Community
  – Social media allows communities to form quickly and communicate effectively. Communities share common interests, such as a love of photography, a political issue or a favorite TV show.

• Connectedness
  – Most kinds of social media thrive on their connectedness, making use of links to other sites, resources and people.
Why Social Media Mining Now?

- Low entry barriers for any internet users
- Need for sharing ideas, collective thinking, debating, writing
- Desire to find new friends, or rediscover the old ones
- People can find information, inspiration, like-minded people, communities and collaborators faster than ever before.
- Coordinating and collaborating
- Gauging group or collective sentiments
- New ideas, services, business models and technologies emerge and evolve amazingly fast
Social Media Offers a New Way of Communication, Participation, and Interactions

• People are connected in a new dimension at an unseen level.
• Participants use social media as an effective tool in social movements and activities, Occupy Wall Street, London Riots, the Arab Spring, and Humanitarian Assistance and Disaster Relief.
• For users, social media is a facilitator for the good, bad, or ugly in a real world (e.g., the recent Assam exodus in August, 2012)
• For researchers, it is a new kind of lab of unprecedented scale for new discoveries.
A New Interdisciplinary Area

• It is propitious time for scientists in social and computer sciences to work together with complementary expertise
  – Social theories distilled from years of research
  – Computer technology and scalable algorithms to deal with Big Data

• New challenges and opportunities for genuine collaborative research

• What can computer or data scientists contribute further in this exciting new development?
What is so Special about Social Media Mining?

• Social Media Mining consist of
  – Social networks
    • Communities
    • Interactions
  – Social media
    • Observational data and lots of them
    • Both content and link information
  – Mining
    • Modeling
    • Learning
    • Prediction
Some High-Level Challenges

- Social media data is a new type of Big Data
  - It differs from conventional (attribute-value) data: linked, noisy, short, unconventional, …

- Social media data is not readily available for scientific investigation
  - Many tasks precede investigation: Data collection, processing, categorization, tracking, and modeling

- Social science theories may not be directly applicable
  - Incessant, self-reported data on relationships now available
  - Ultimately, there are fundamental questions such as: how similar the two worlds are, how to measure the findings, …
Some Specific Challenges

• Everyone can publish
  – Challenge 1: How can I be heard?

• There exist a multiplicity of online sources
  – Challenge 2: Which one should I use?

• People participate in activities on Web
  – Challenge 3: How can one improve his/her experience and create relevant business opportunities?
Challenge 1: How can I be heard

• Everybody can speak up, but people only listen to a few
  – Information is cheap, but attention is costly.
  – If you were one of the few, you could at least make living without any “traditional” job

• Sometimes, a real treasure might be buried forever
  – How to discover hidden treasures
  – If that is difficult, how to make us first visible
  – Study how a star develops, shines, and what others do (influence, diffusion, and cascading)
Challenge 2: Which source should I use

- Information overloading
  - Abundance is good, but too many choices can mean no choice (any déjà vu?)

- Information fusion or integration
  - Filtering irrelevant and/or redundant information
  - Discovering complementary information
  - Comparison shopping

- Trust and reliability
  - How to discern the real from the fake (e.g., rumors)
  - Spam (Any of you experience it in FB recently?)
Challenge 3: Improve user experience

- A business perspective – how to help generate profit for a business or make `me’ rich
  - Better experience means (or leading to ) more traffic or more purchases
- Understanding
  - Who are they? What are their sentiments?
  - Who influence them? How to engage their influencers?
  - Who are moving away from my site (migration)?
- “Data, data, data”, and “big data”
  - Data collection and analysis
  - Often it is about lots of little data with links (an egocentric perspective)
Some Examples of Research Problems

• Importance of weak ties
  – How to define them? How do they impact on influence propagation and information diffusion?

• Are our “friends” really friends?
  – Should I trust everyone? Whom should I trust and on what? Is it achievable?

• Which social media data is good for which social computing tasks?
  – A typical Big-Data problem: need to be selective
• What is the next big impact of computer science?
  – Bioinformatics manifests the significant and lasting impact of CS on biology
  – Pervasive use of Web and Internet fosters cultural and behavioral changes
    • Friendship networks
    • Citizen journalism and Blogosphere
    • Folksonomy (digg, reddit)
  – Vigorous activities in a virtual world and in the boundary of virtual and physical worlds
  – An unprecedented laboratory for research
Developing Theories

• Social science theories
  – From a dozen of people to over thousands
  – Searching for applicable, adaptable, or new theories

• Computer science theories
  – New data types, and new data sources
  – Computing with collective intelligence

• Computational social theories
  – Using a computational “telescope” to “look beyond what we can see” about human and social dynamics
  – Simulation to understand (e.g., Thomas Schelling’s experiment)
Advancing Applications

• Using social media as a lens to understand human behavior
  – When and where can we use social media patterns to predict real-world patterns

• Facilitating humanitarian assistance and disaster relief
  – Tracking, preparing and coordinating for HADR
  – Knowing where/who “you” are, whenever needed

• Improving crowdsourcing capabilities
  – Expansive, but noisy and/or flimsy
  – How can we get relevant and useful information timely
Research with Social Media

- Community Analysis
- Provenance and Trust
- Migration
- Influence & Diffusion
- Location based Social-network Analysis
- User Vulnerability
- Behavior Analysis
- Research Issues
- Methodology & Tools
- ACT
- Tweet & Blog Tracker
- Linked Feature Selection

Data Mining and Machine Learning Lab
• Where are the problems?
  – Where do we want to gain insights and deep understanding?
• Are they interesting, relevant, and challenging?
• Is it feasible to solve them?
• Can we find better solutions?

• These questions and others challenge us to *move the research frontier forward* through interdisciplinary collaboration.
• Now we present some specific examples to illustrate social media mining
Users share a large amount of personal information

Malicious individuals or groups may take advantage of openness and naivety of users by exploring vulnerability.

Such exploration can cause physical and emotional distress to users, sometimes with dire consequences.

Many a time, a user is unaware that he could pose a threat to his friends.
Challenges

• How to measure user vulnerability?
• How to define vulnerable users and find them?
• Can we make users less vulnerable without suggesting fundamental changes to its architecture?
• What measures can one take for protection?
• What is the impact of new friends on the vulnerability of a user?
Measuring Individual User’s Action

- Index-based on visibility
  - Individual attributes
  - Community attributes

- How visible a user is?
  - We estimate the visibility score of a user using both attributes

Gundecha et. al. “Exploiting Vulnerability to Secure User Privacy on a Social Networking Site”. (SIGKDD 2011)
User Vulnerability

- A user vulnerability depends on actions of individual user and his friends
- We estimate the user vulnerability score using the visibility score of individual user as well as his friends
Defining a Vulnerable Friend

A vulnerable friend of a user is defined as a friend whose removal or unfriending will lower the user vulnerability score of a user.
K-Vulnerable Friends

- A k-vulnerable friend of a user is a set of k friends whose removal or unfriending will lower the user vulnerability score of a user.
Performance Comparisons for Unfriending

- (a) Most Vulnerable Friend
- (b) Least V-Friend
- (c) Random Friend
- (d) Max V-Friend
Vulnerable Friends with High Social Utility

- What if the vulnerable friends are of some importance such as girlfriend, boyfriend or influential friends?
- Can we reduce vulnerability without sacrificing social utility?
## Social Utility Measures

<table>
<thead>
<tr>
<th>Type</th>
<th>Social utility measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Each vulnerable friend</td>
<td>Degree</td>
</tr>
<tr>
<td></td>
<td>Local clustering coefficient</td>
</tr>
<tr>
<td></td>
<td>Number of closed triads</td>
</tr>
<tr>
<td></td>
<td>Number of open triads</td>
</tr>
<tr>
<td></td>
<td>Number of posts</td>
</tr>
<tr>
<td></td>
<td>Number of responses</td>
</tr>
<tr>
<td></td>
<td>Popularity / Influence index</td>
</tr>
<tr>
<td>Relationship between user and each vulnerable friend</td>
<td>Priority (social tie strength)</td>
</tr>
<tr>
<td></td>
<td>Number of common friends</td>
</tr>
<tr>
<td></td>
<td>Number of uncommon friends (structural holes)</td>
</tr>
<tr>
<td></td>
<td>Number of interactions</td>
</tr>
<tr>
<td></td>
<td>Homophily (similarity)</td>
</tr>
</tbody>
</table>
Vulnerability Minimization

• Problem: Given the goal to maintain low vulnerability for a user, is there a feasible solution that can retain high social utility value?

• Vulnerability minimization problem can be equivalently stated as finding the set of vulnerable nodes which maximizes vulnerability.

• We formulate this as optimization problem and provide an approximation algorithm with a proven bound.
Unfriending at Most K-vulnerable Users while Retaining Degree Utility (at least 90%)

(a) Unfriend 1

(b) Unfriend 2

(c) Unfriend 10

(d) Unfriend 50
Summary

• Not only a user but also his friends should have responsibility to protect their privacy.
• We present a measure to estimate user vulnerability.
• Secure users needs to be more careful about the new friend requests.
• It is possible to reduce vulnerability without sacrificing social utility.
How Vulnerable Are Your Facebook Friends

By Karyl Scott / August 24, 2011 1:30 PM / 1 Comment

Even as Facebook moves to create a social network’s 750 million users, some researchers are proposing even greater concerns for social networks that place little trust in their users.

The researchers are presenting their work at the Conference on Knowledge Exploration and Applications. They’ve created an index that lets users determine the most likely people they should remove from their networks based on their privacy settings, behavior and the interactions between their accounts.

For example, around 80% of users are happy to disclose their gender, but less than 1% share their home address. That suggests people publicizing their address aren’t particularly privacy-conscious and you might want to avoid them.

The solution, says Pritam Gundecha, a computer science PhD student at Arizona State University and one of the authors of the study, is to unfriend those with questionable behaviors or friends who have not set their privacy controls to acceptable levels.

If you don’t wish to unfriend someone, you can make them aware of their vulnerabilities and to ask them to address them, says Gundecha, whose studies focus on social media security.

How to determine which of your Facebook friends to get rid of

By Jacob Aron / Magazine issue 2827. Subscribe and save

WORRIED about loose-tongued friends sharing your private details with the world? Calling the least discreet members of your social network will help you feel more secure, but it’s not a perfect solution. What if your best friend is an offender?

Google’s social networking site, Google+, had been running for less than a week when it turned out there was nothing to stop your friends “resharing” posts with the entire internet. Google now lets users disable reshares, but the problem is indicative of how little control you have over what your friends do.

For example, around 80% of users are happy to disclose their gender, but less than 1% share their home address. That suggests people publicizing their address aren’t particularly privacy-conscious and you might want to avoid them.

Pritam Gundecha at Arizona State University in Tempe has a technique for working out which friends are most likely to leak private information so you can remove them, if you choose. Gundecha examined the relative importance of data 2 million Facebook users elect to share with the world and calculated the privacy risks friends pose to each other.

For example, around 80% of users are happy to disclose their gender, but less than 1% share their home address. That suggests people publicizing their address aren’t particularly privacy-conscious and you might want to avoid them.

The solution, says Pritam Gundecha, a computer science PhD student at Arizona State University and one of the authors of the study, is to unfriend those with questionable behaviors or friends who have not set their privacy controls to acceptable levels.

If you don’t wish to unfriend someone, you can make them aware of their vulnerabilities and to ask them to address them, says Gundecha, whose studies focus on social media security.

Toronto Star

(Canadian)

ReadWriteWeb

(Syndicated on NY Times)

New Scientist

(UK)
Location-Based Social Networks (LBSNs)

- Foursquare, Gowalla (Facebook)
- Essential Contents Provided by LBSNs
  - **Check-in** history with **time stamps**
  - **Social Networks** derived from check-in locations
  - **Interdependency** of social networks and locations
- Significance of “Check-in” behavior
  - Reflect the interactions between a user and the world
  - Bridge the gap between the real world and online social networks.
Social-Historical Effect on Predicting Online Check-ins

Historical Ties

T1

T2

T3

T4

Social Ties

Friend A

Friend B

User

Friend C

Friend D
Challenges in Analyzing User’s Historical Ties

- Power-law distribution

Short-Term Effect of Historical Ties

- The historical tie strength decreases over time.
- The historical ties of the previous check-ins at airport, shuttle stop, hotel and restaurant have different strengths to the latest check-in of drinking coffee.
Modeling User’s Historical Ties

- Correspondences between language and LBSN modeling

<table>
<thead>
<tr>
<th>Language Modeling</th>
<th>LBSN Modeling</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corpus</td>
<td>Check-in collection</td>
</tr>
<tr>
<td>Document Structure</td>
<td>Individual check-ins</td>
</tr>
<tr>
<td>Paragraph</td>
<td>Check-in Structure</td>
</tr>
<tr>
<td>Sentence</td>
<td>Monthly check-in sequence</td>
</tr>
<tr>
<td>Phrase</td>
<td>Weekly check-in sequence</td>
</tr>
<tr>
<td>Word</td>
<td>Daily check-in sequence</td>
</tr>
<tr>
<td></td>
<td>Check-in location</td>
</tr>
</tbody>
</table>

- Power-law distribution
- Short Term Effect

HPY (Hierarchical Pitman-Yor) Language Model
Modeling User’s Social Ties

- **Friends Check-in History**
  - Friends have higher check-in similarity than strangers.

<table>
<thead>
<tr>
<th></th>
<th>Common check-ins</th>
</tr>
</thead>
<tbody>
<tr>
<td>between friends</td>
<td>11.8306</td>
</tr>
<tr>
<td>between strangers</td>
<td>4.3226</td>
</tr>
</tbody>
</table>

- **Friends Check-in History**
- **User Similarity**

\[
\text{HPY (Hierarchical Pitman-Yor) Language Model}
\]
Experiment Results for Location Prediction

- **MFC** (Most Frequent Check-in Model)
- **MFT** (Most Frequent Time Model)
- **Order-1** (Order-1 Markov Model)
- **Order-2** (Order-2 Markov Model)
- **HM** (Historical Model)
- **SHM** (Social-Historical Model)
Relationship between Social-historical Model
Trust in a Connected World

- Role of trust is critical in some online communities such as product review sites and e-commerce sites.
- Users in product review sites (like Epinion) provide product reviews as well as trust networks.
- Trust networks are widely exploited in review quality prediction, and recommender systems to improve accuracy.
- Most of the current research assumes single and homogenous trust relationships between users.
Multi-Faceted Trust

- A Multi-Faceted Trust Example from Epinions

Methodology

- **Representation**
  - Adjacency Matrix
  - Tensor

- **Trust Strength Inference**
  - Rating Similarity and Trust
  - Rating Prediction

\[ \hat{R}(u, i) = \alpha \left( \frac{\sum_k PF(i, k)\mu(k) + B(u, k))}{\sum_k PF(i, k)} + c(i) \right) + \frac{\sum_{k=1}^{K} \sum_{v \in N(u, i)} PF(i, k)A(v, u, k)R(v, i)}{\sum_{k=1}^{K} \sum_{v \in N(u, i)} PF(i, k)A(v, u, k)} \]
Findings

- Heterogeneous Trust Links
  - Pairs of Reciprocal Links
  - Transitive Trust Relationships
  - Co-Citation
Applications of mTrust

• With these findings, mTrust can be applied to some online tasks.
  – Improving Rating Prediction
  – Enabling Facet-Sensitive Ranking
  – Making Status Theory (an important social theory) applicable to reciprocal links
Real-World Applications

• **ACT (ASU Coordination Tracker)**
  - An event response coordination system
  - Facilitate multiorganization response (military, governments, NGOs, etc.) to an event such as a disaster
  - Provide relief organizations the means for better collaboration and coordination during a crisis.

• **TweetTracker**
  - a Twitter-based analytic and visualization tool.
  - Help relief organizations to acquire situational awareness during disasters and emergencies to aid disaster relief efforts.
Conclusions

• Existing data mining techniques are not adequate to handle social media data
  – Social media data is vast, noisy, distributed, unstructured and dynamic
• Social media mining is a burgeoning area where more challenges are present than solutions
• Interdisciplinary research will surely help us develop theories and advance applications
Additional Information

- Modeling and Data Mining in Blogosphere (2009)
- Community Detection and Mining (2010)
- Provenance Data in Social Media, to appear

**SBP Conference Series**

- SBP08–12 Proceedings
- SBP13
  April, D.C.

SDM 2012
Anaheim, CA
Acknowledgments: We thank DMML members for their comments and our collaborators for their wisdom and contributions. This work is, in part, supported by ONR, ARO, and NSF.

We appreciate the invitation by INFORMS for an opportunity to present this tutorial.