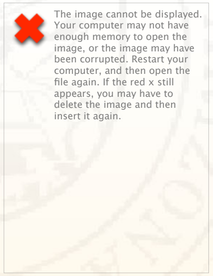
Some Computational Challenges in Mining Social Media

Huan Liu

August 26th, 2013 Niagara Falls, Canada

Traditional Media and Data

Broadcast Media
One-to-Many



Communication Media
One-to-One

Traditional Data

Arizona State University

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
  - Collaborative environment, and wisdom of the crowd
  - Many small groups (the long tail phenomenon)
  - Attention is expensive

A Big Variety of Social Media

- Networking
- Sharing
- Publishing
- Gaming
- Discussing
- Location
- Marketing

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 New Data-Mining Challenges in SM

- Evaluation Dilemma
  - Evaluation without conventional test data, but how?
- Big-Data Paradox
  - Often we get a small sample of (still big) data. How can we ensure if the data can offer credible findings?
- Noise-Removal Fallacy
  - How do we remove noise without losing too much?
- Finding Needles in a Changing Haystack
  - Social media is full of “everything”. Where can we find the relevant information we need **fast**?

Challenge 1: Evaluation Dilemma

- In conventional data mining, training and test datasets are used to validate findings and compare performance.
- Without training-test data and with the need to evaluate, how can we do it?
  - User study, Amazon Mechanical Turk, …
  - Are they scalable, reproducible, or applicable?
- We need to explore new ways of evaluation.

Understanding User Migration Patterns in Social Media

Joint work with Shamanth Kumar and Reza Zafarani

AAAI 2011, San Francisco, CA

Limited Resources with Increasing # of Sources

- Hundreds of social media sites and many more appearing
- We all have only limited resources (time, energy, …)
  - Cannot be active on all the sites
  - Must choose sites to participate
- Migration between sites may be inevitable

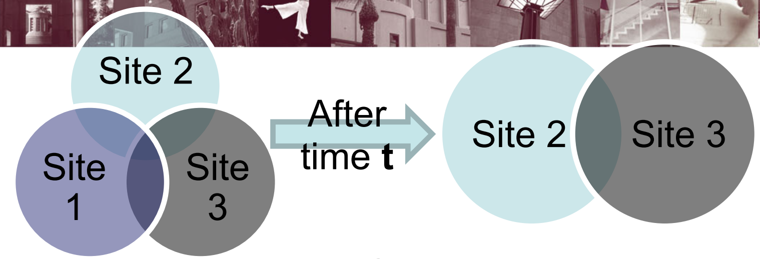
Why Do We Care about Migration?

- Users are a primary source of revenue
  - Ads, Recommendations, Brand loyalty
- New social media sites need to attract new users to expand their user base
- Existing sites need to retain their users by migration prevention
- Competition for attention entails the understanding of migration patterns

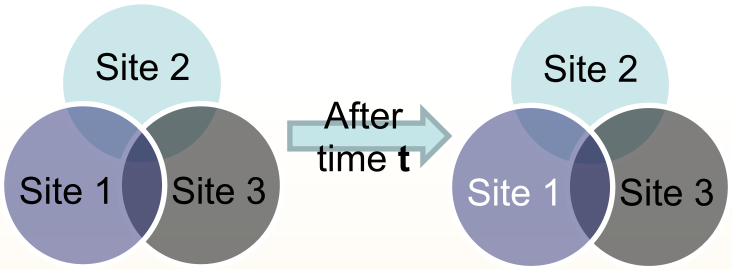
Migration in Social Media

- What is migration?
  - Migration can be described as the movement of users away from one location toward another, either due to necessity, or attraction to the new environment.
- Migration in social media can be of two types
  - Site migration
  - Attention migration

Types of Migration



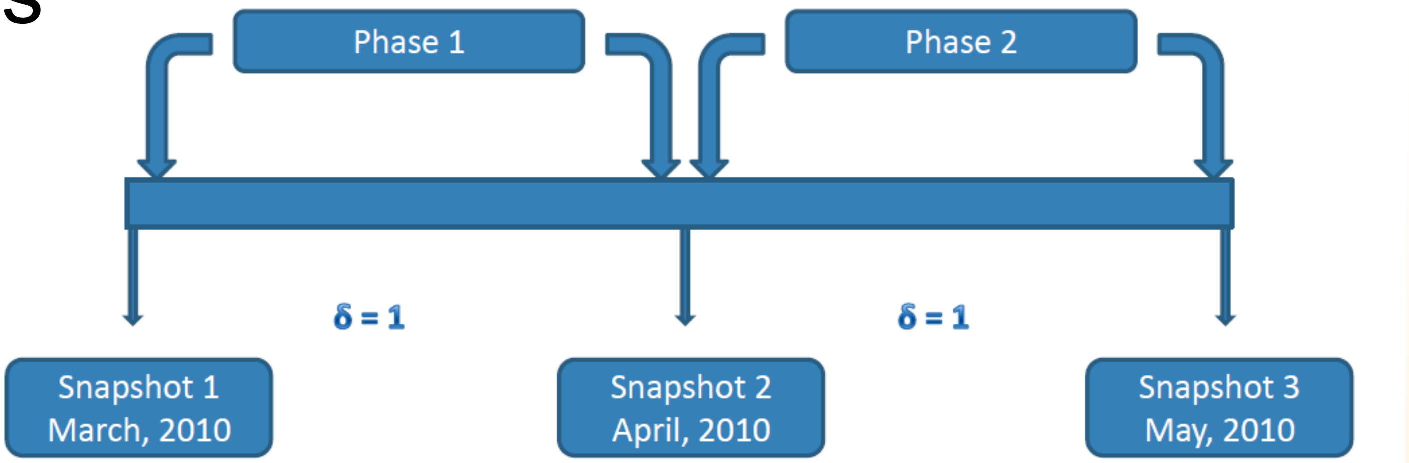
- Site Migration
  - Users leave a site by profile deletion or profile removal
  - Difficult to convince a user who has deleted his account to return
  - Hard to study these users cross site because we need their registration information



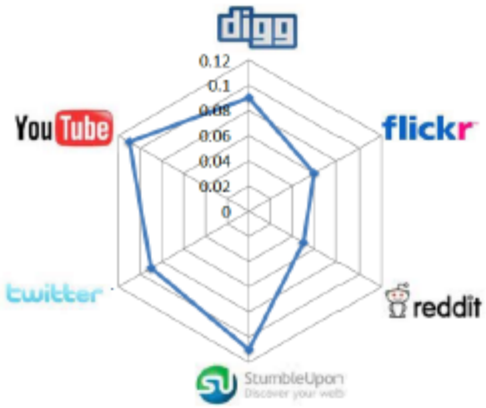
- **Attention Migration**
  - Users can become inactive on a site
  - A harbinger for site migration
  - Can be studied to prevent site migration by understanding migration patterns
  - Can be detected by observing *user activities* across sites

Obtaining User Migration Patterns

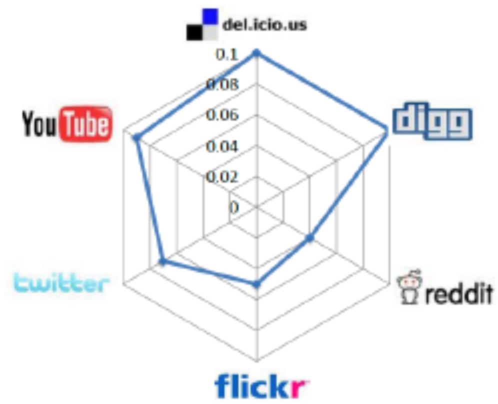
- Goal: Identifying trends of attention migration of users across the two phases of the collected data.
- Process



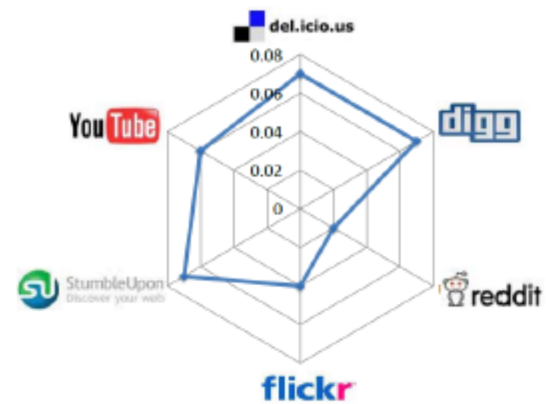
Patterns from Observation



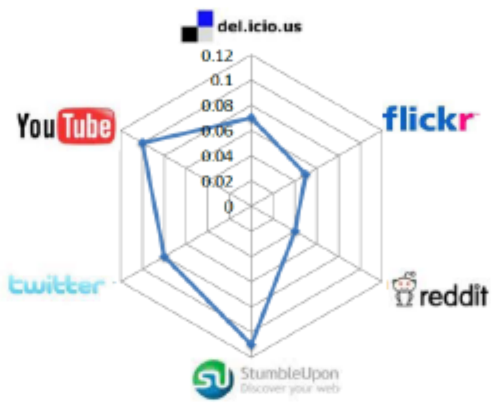
(a) Delicious



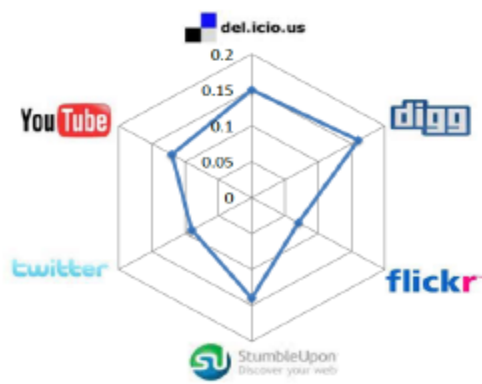
(e) StumbleUpon



(f) Twitter



(b) Digg



(d) Reddit

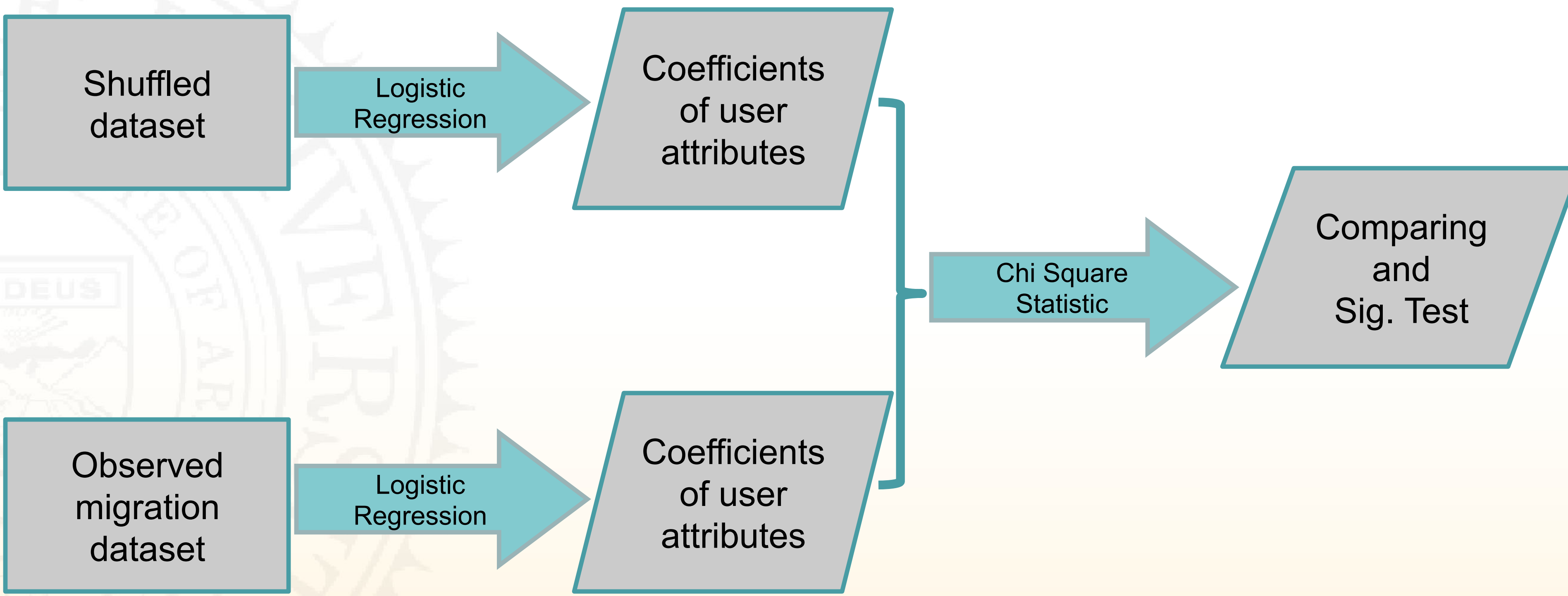
Facing an Evaluation Dilemma

- Important to know if they are valid or not
  - If yes, we investigate further how we use the patterns to: prevention or promotion.
  - If not, why not? And what can we do?
- We would like to evaluate migration patterns, but without ground truth
- How?
  - User study or AMT?

Evaluating Patterns' Validity

- One way is to verify if these patterns are fortuitous
- Null Hypothesis: *Migration of individuals is a **random** process*
  - Generating another similar dataset for comparison
    - Potential migrating population includes overlapping users from Phase 1 and Phase 2
    - Shuffled datasets are generated by picking random active users from the potential migrating population
    - The number of random users selected for each dataset is the same as the real migrating population

A Significance Test



Evaluation Results

- Significant differences observed in StumbleUpon, Twitter, and YouTube
- Patterns from other sites are not statistically significant. Potential cause:
  - Insufficient Data?

Table 2: χ^2 test results on the observed and shuffled data

Site	Observed Coefficients			Shuffled Coefficients			p-value	Statistical Significance
	N	A	R	N	A	R		
Delicious	0.2858	0.4585	-	0.6029	0.5921	-	0.65	Not significant
Digg	0.4796	0.8066	-	0.52	0.5340	-	0.70	Not significant
Flickr	1	1	0.9797	0.2922	0.2759	0.4982	0.13	Not significant
Reddit	0.5385	0.6065	-	0.4846	0.6410	-	0.92	Not significant
StumbleUpon	1	1	-	0.4191	0.2059	-	0.0492	Significant
Twitter	0.5215	1	0.5335	0.2811	0.0365	0.4009	0.0001	Extremely significant
YouTube	0	1	0.1644	0.7219	0.0040	0.4835	0.0001	Extremely significant

Summary

- Mitigating or promoting migration by targeting high net-worth individuals
  - Identifying users with high value to the network, e.g., high network activity, user activity, and external exposure
- Social media migration is first studied in this work
- Migration patterns can be evaluated without test data

Challenge 2: Big-Data Paradox

- What is Big Data?
  - 3Vs, 4Vs, or 5Vs …
- Is Social Media Data big?
  - Yes, it is obviously so (e.g., FB and Twitter)
  - But, we are often limited by source APIs (e.g., Twitter Streaming API offers 1% data)
- What can we do facing the cold reality?
  - It depends on if you're rich, famous, lazy, or curious

Is the Sample Good Enough? Comparing Data from Twitter's Streaming API and Data from Twitter's Firehose

Joint Work with Fred Morstatter, Jürgen Pfeffer, and Kathleen Carley

AAAI ICWSM2013, Boston, MA

Big-Data Problems

- 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.

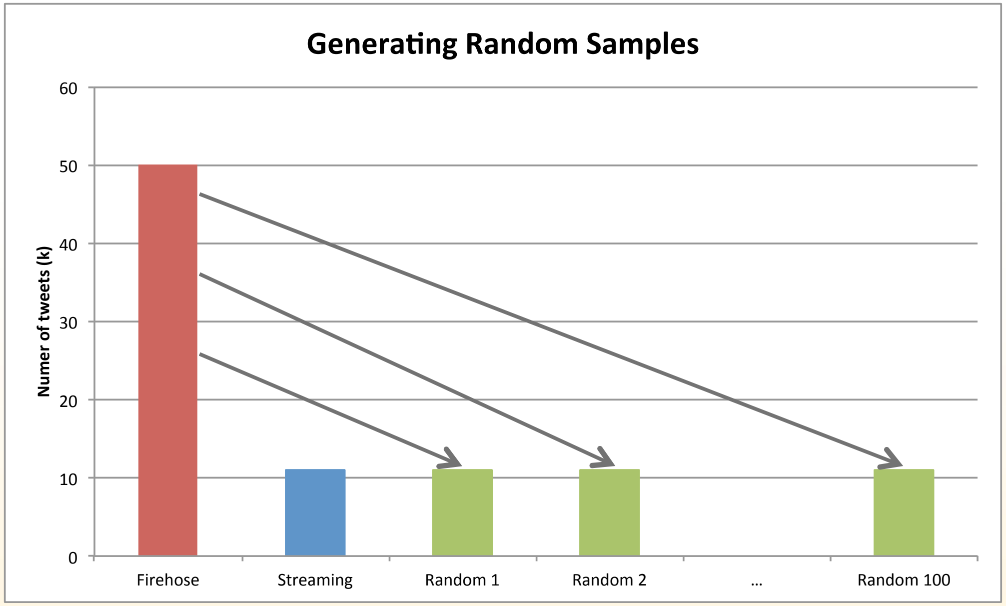
Network Measures	Geographic Distributions
• Found ~50% of the top tweeters by different centrality measures. • Graph-level measures give similar results between the two datasets.	• Streaming data gets >90% of the geotagged tweets. • Consequently, the distribution of tweets by continent is very similar.

How are These Results?

- Accuracy of Streaming API varies with analysis the researcher wants to perform.
- 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?

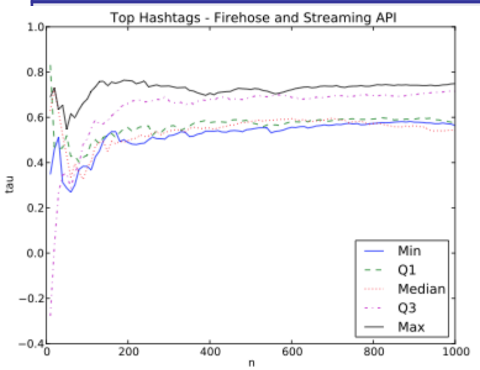
Probing Further

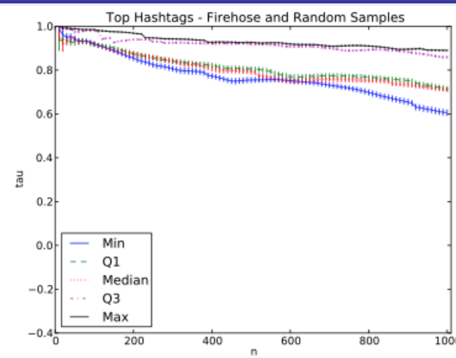
- Aggregate data by day
- Select 5 days with different levels of coverage
- Create random samples from the Firehose data to compare against the Streaming API

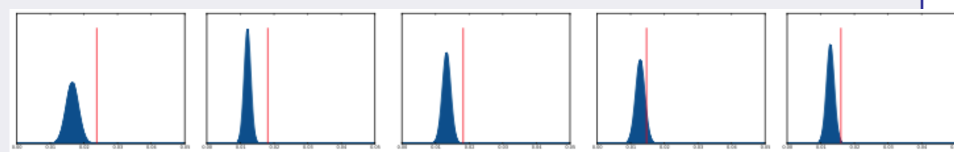


Comparative Results

Top Hashtags	Topic Extraction







- No correlation between Streaming and Firehose data.
- **Not so in random samples**

- Topics are close to those found in the Firehose.
- **Topics extracted with the random data are significantly better.**

Summary

- Streaming API data can be biased in some facets.
- Our results were obtained with the help of Firehose.
- Without Firehose data, challenges are to figure out which facets have biases, and how to compensate them in search of credible mining results

Challenge 3: Noise-Removal Fallacy

- A common complaint: "99% Twitter data is useless".
  - "Had eggs, sunny-side-up, this morning"
  - Can we remove the 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?

Feature Selection with Linked Data in Social Media

Joint Work with Jiliang Tang

SDM2012 and KDD2012

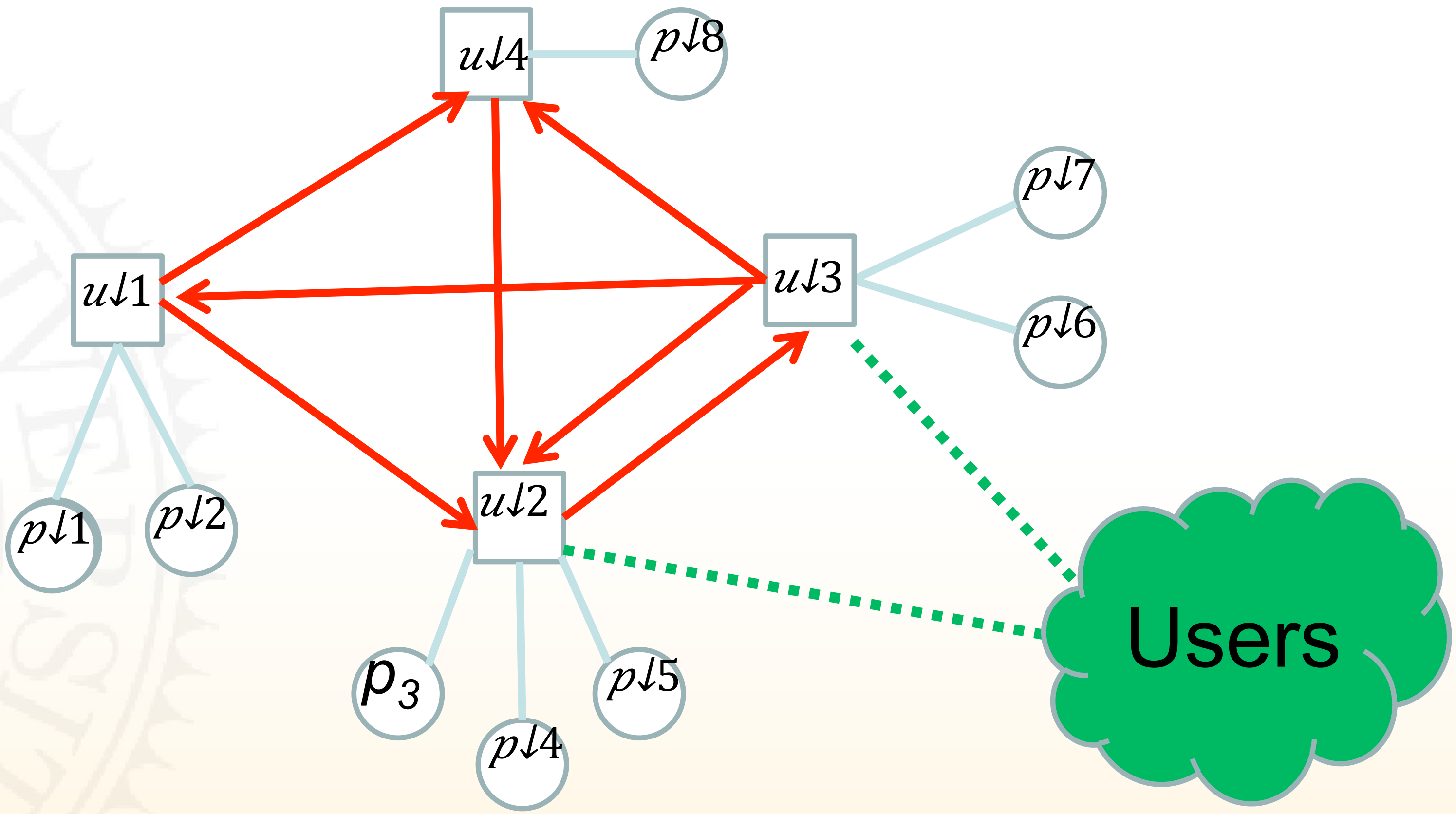
Social Media Data

- 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

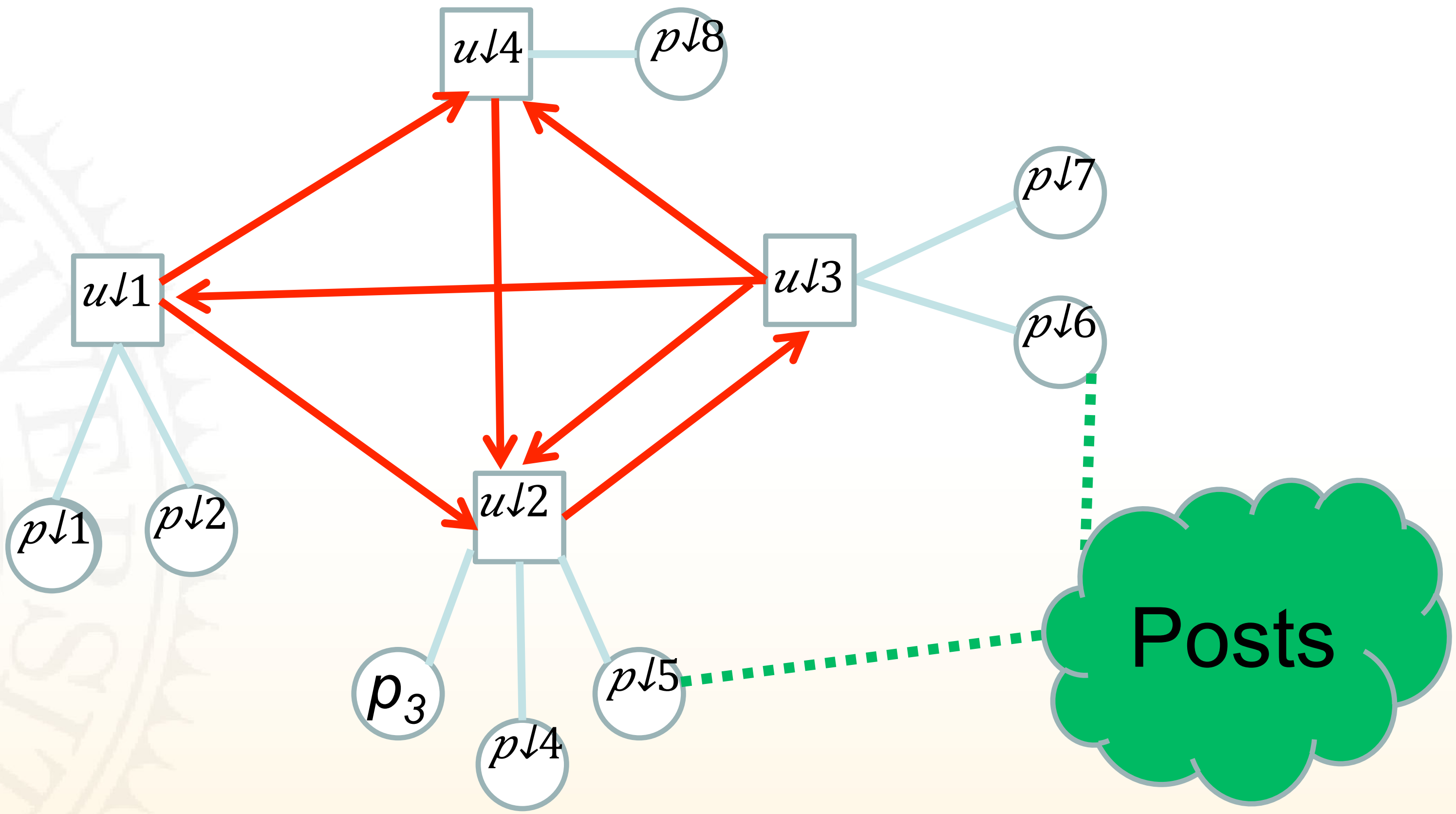
Feature Selection of Social Data

- 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

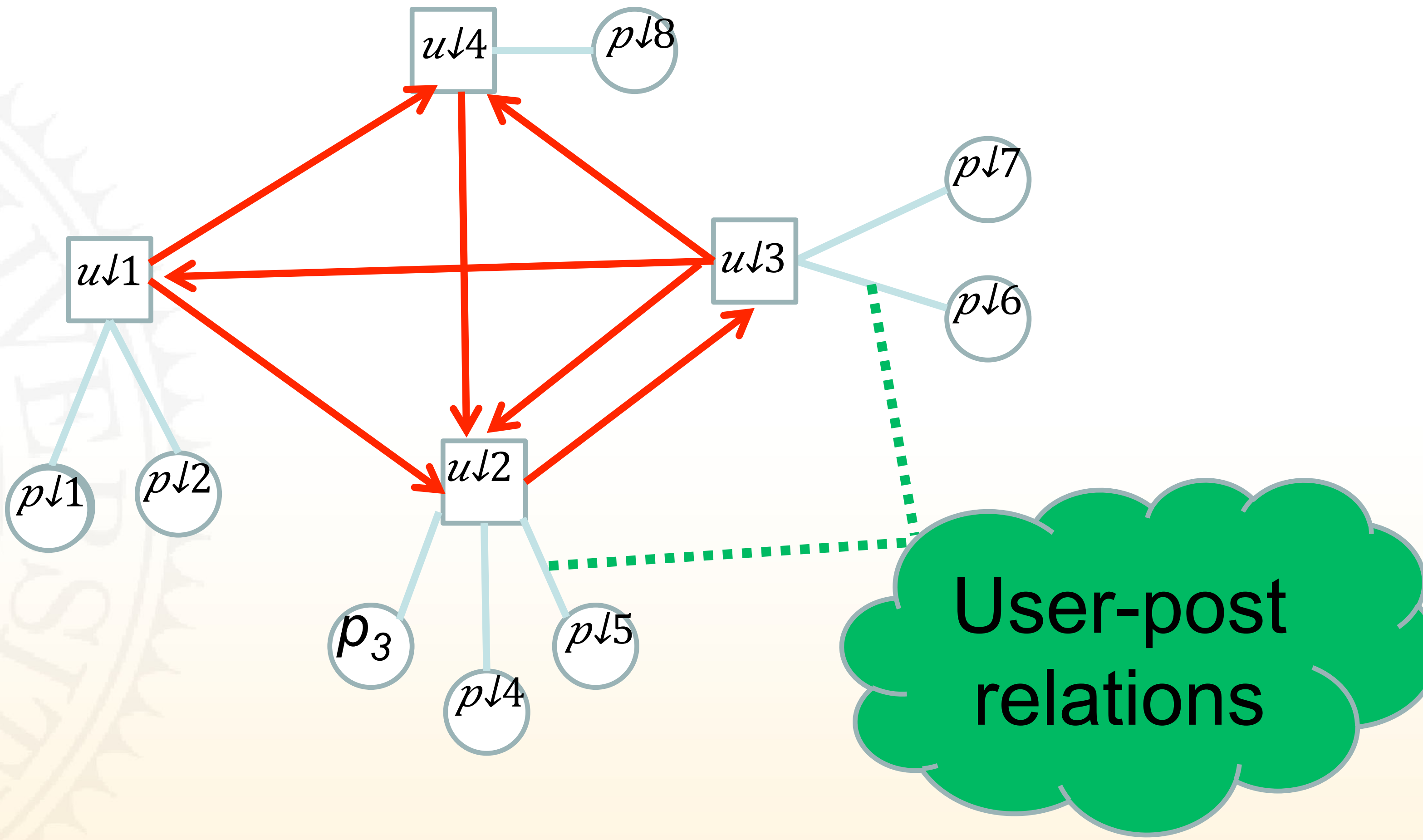
An Example of Social Media Data



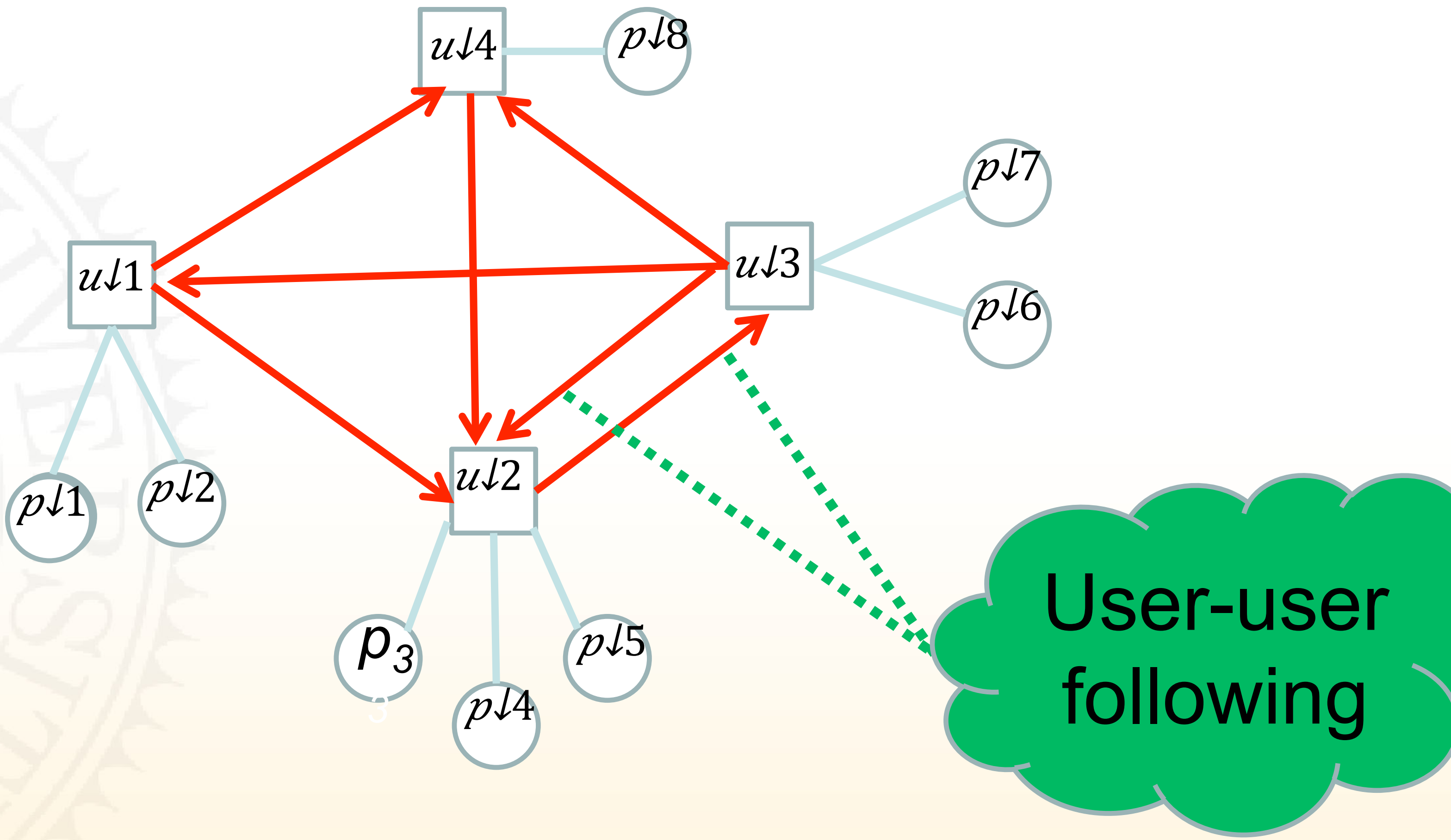
An Example of Social Media Data



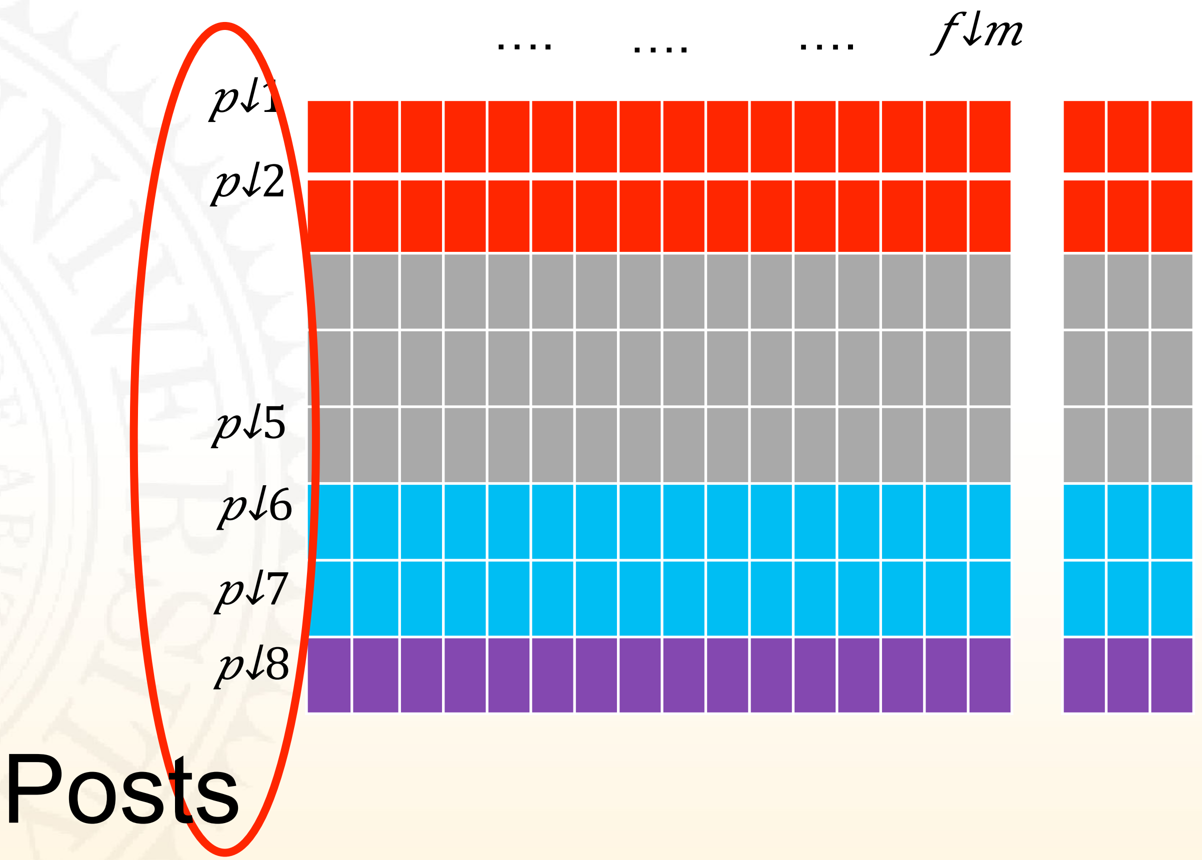
An Example of Social Media Data



An Example of Social Media Data

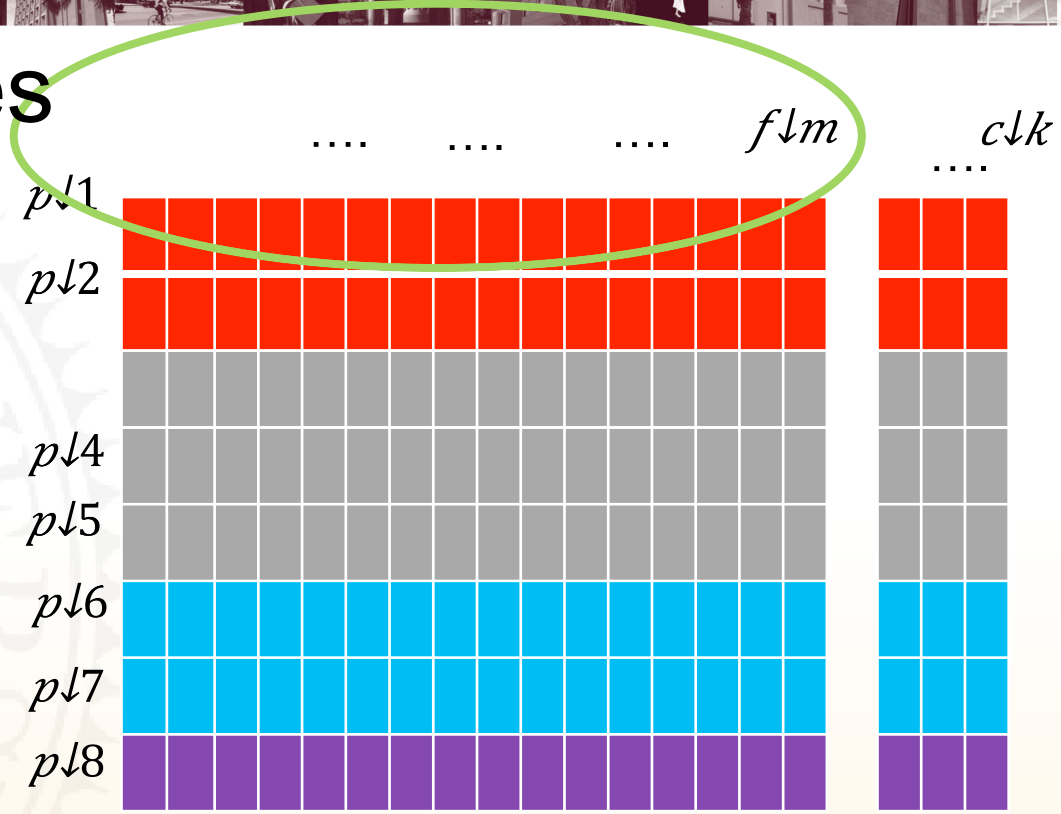


Representation for Attribute-Value Data

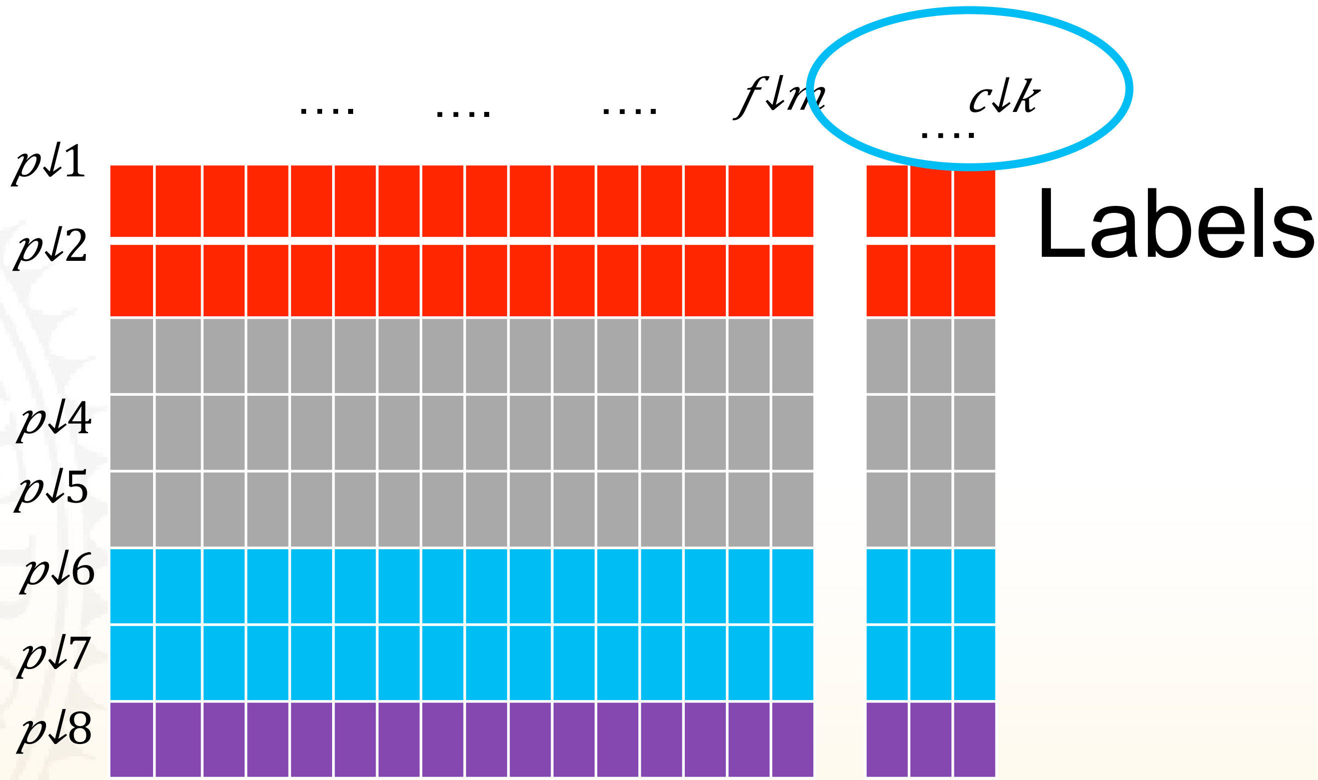


Representation for Attribute Value Data

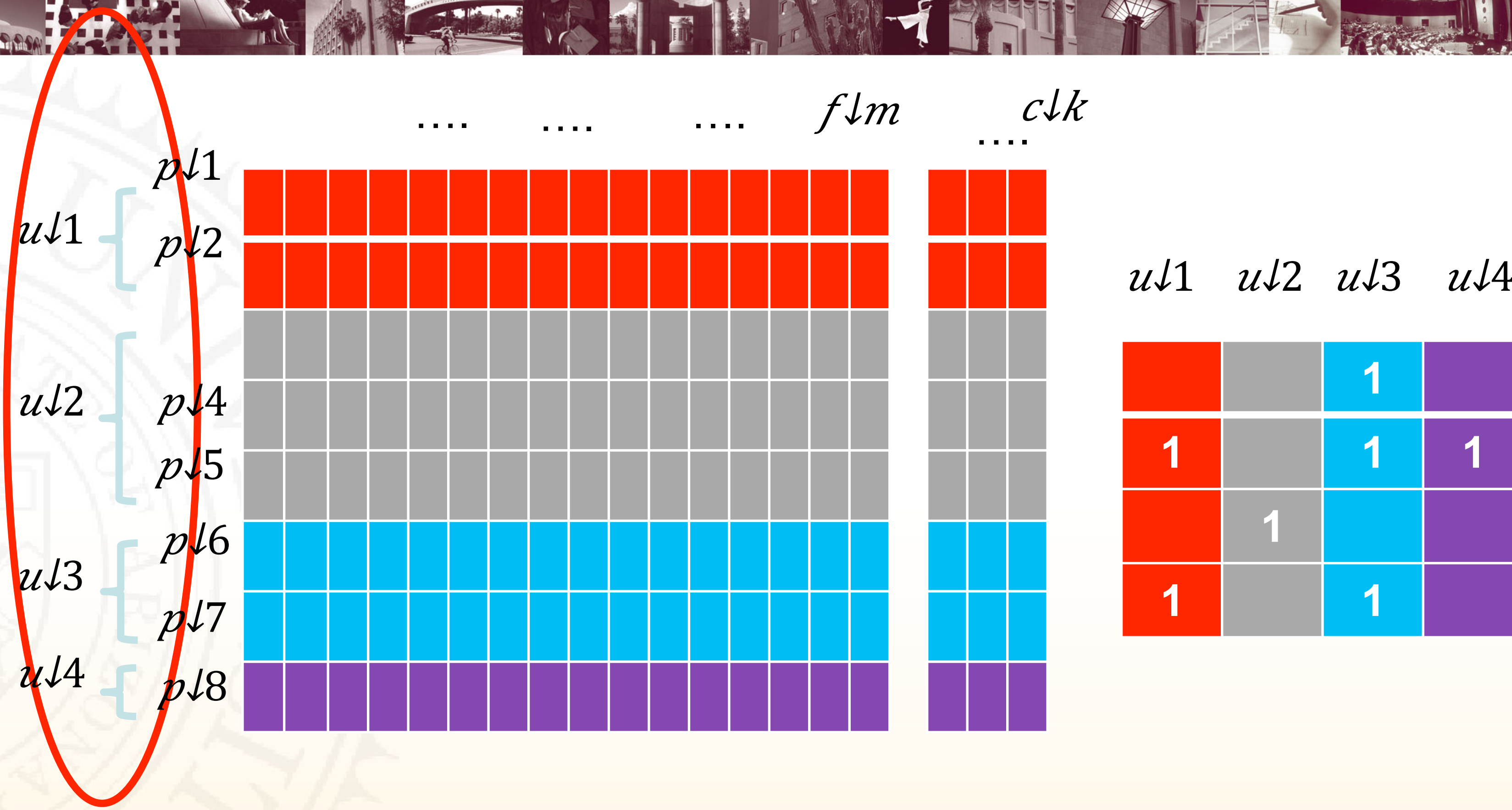
Features



Representation for Attribute Value Data

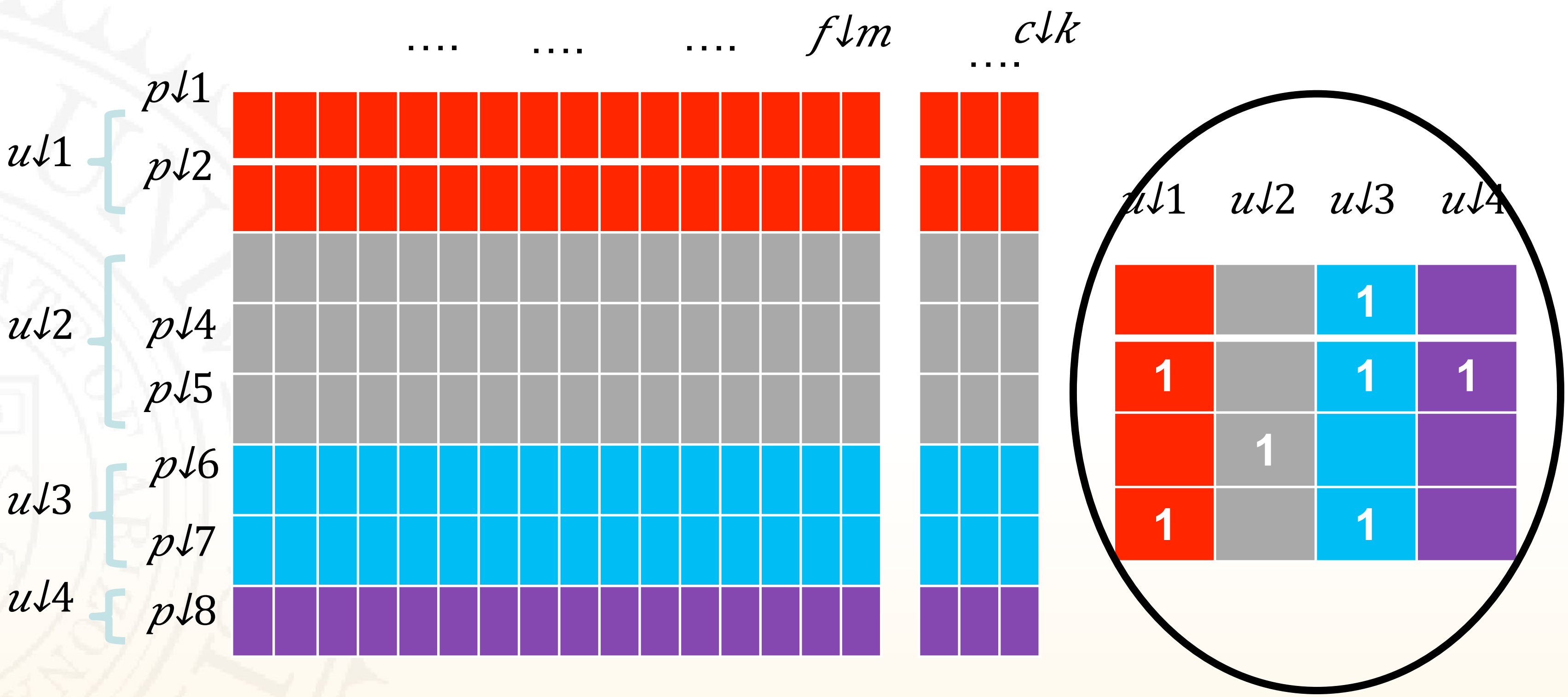


Representation for Social Media Data



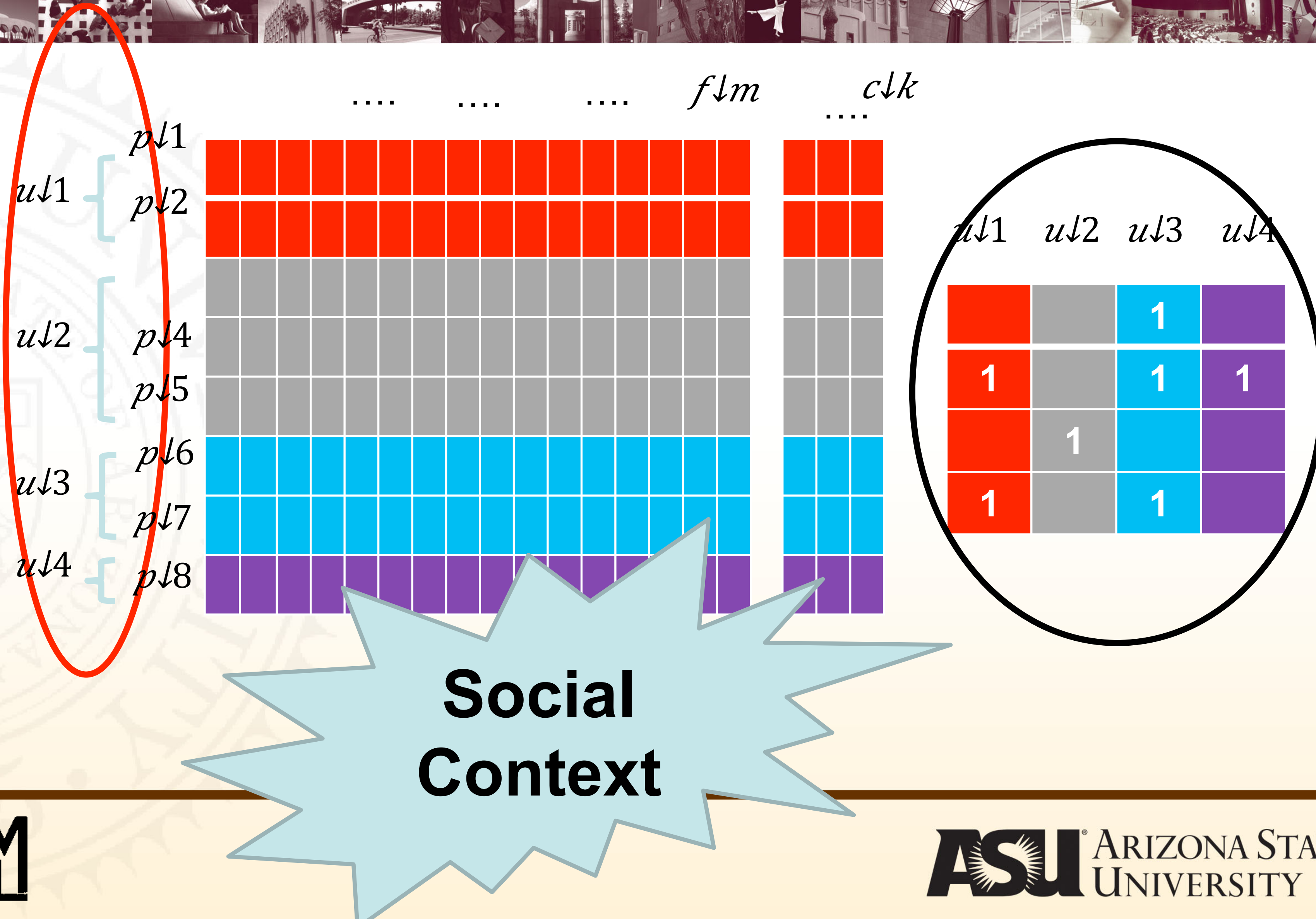
User-post relations

Representation for Social Media Data



User-user relations

Representation for Social Media Data



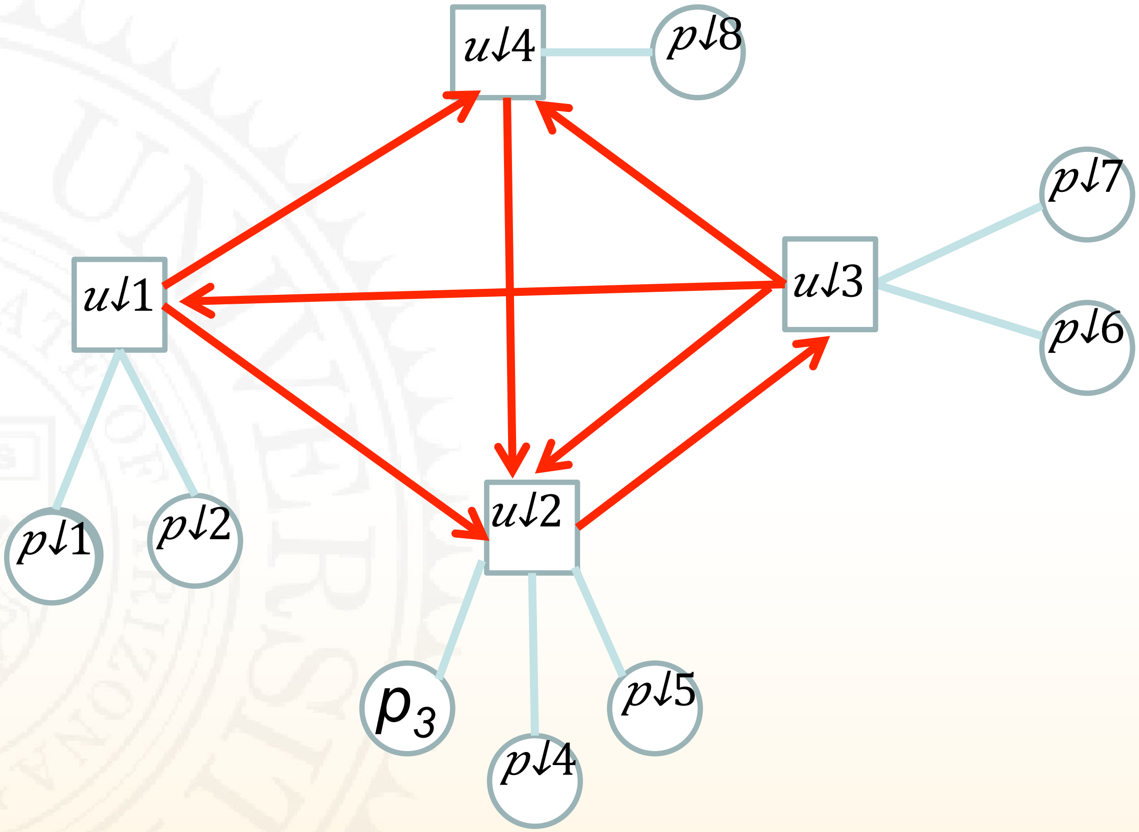
Problem Statement

- 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 $\boldsymbol{k}$ most relevant features from $\boldsymbol{m}$ 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

Modeling CoFollowing Relation

- Users' topic interests

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

- Two co-following users have similar interested topics

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

DMML

ARIZONA STATE UNIVERSITY

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.
- An unsupervised method is showcased in our KDD12 paper following social correlation theories

Challenge 4: Needles in a Changing Haystack

- Social media data is a much messier haystack with *many* hidden, useful *needles*
- With so much and so noisy data, how can we find relevant information for time-sensitive, critical events?
- We need new ways of filtering data for valuable information for various purposes

Whom Should I Follow? Identifying Relevant Users During Crises

Joint Work with Shamanth Kumar, Fred Morstatter, and Reza Zafarani

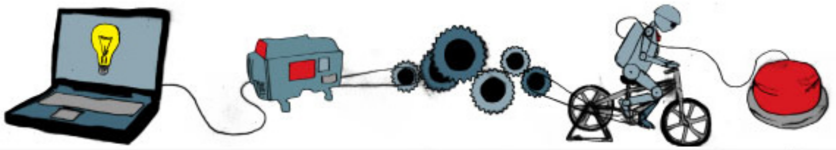
ACM HT2013, Paris, France

Quickly Identifying Relevant Users during Crises

- Motivation
  - Twitter is playing a prominent role in time-sensitive critical events (e.g., the Arab Spring).
  - It offers a unique lens for information collection.
- How can we gain fast access to relevant and useful information during crises, while the data is
  - High-volume
  - High-velocity
  - Noisy
  - Widely spread

One Per Cent

Taking the sweat out of technology



How to find the right Twitter user in a crisis

12:46 20 March 2013

Twitter | social media

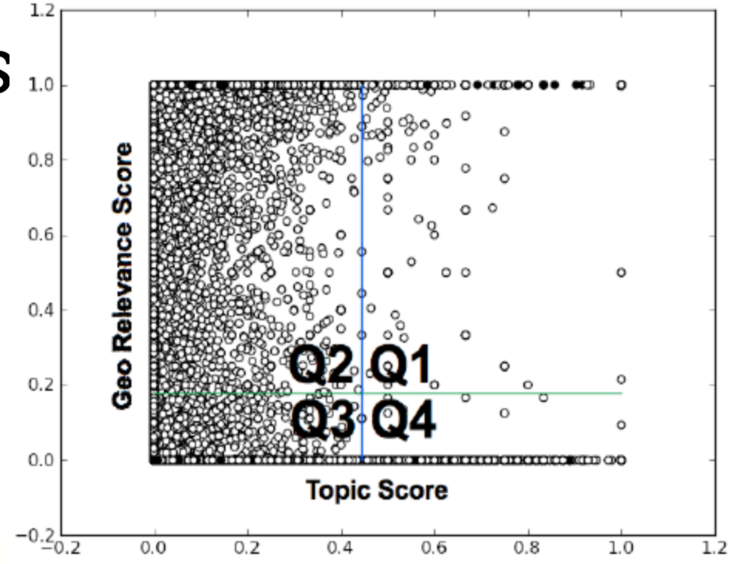
Hal Hodson, technology reporter

(Image: AFP/Getty)

Honing your Twitter feed can be a chore at the best of times, but when you want the latest information during a natural disaster or national uprising, things get a lot harder.

An Approach

- Using *topics* and *user locations* we can classify users into
  - *Q1 (Information Leaders)*
  - Q2 (Topic Ignorant)
  - Q3 (Apathetic)
  - *Q4 (Sympathizers)*



- Q1 users talk about specific topics
- Q4 users talk about similar topics but with different intensity

Our Findings

- Q1 users generate more information than a random set of users
- Q1 users generate high-quality information
- Q1 users post more tweets than users with followers
- Visualized topics show that Q1 are more focused

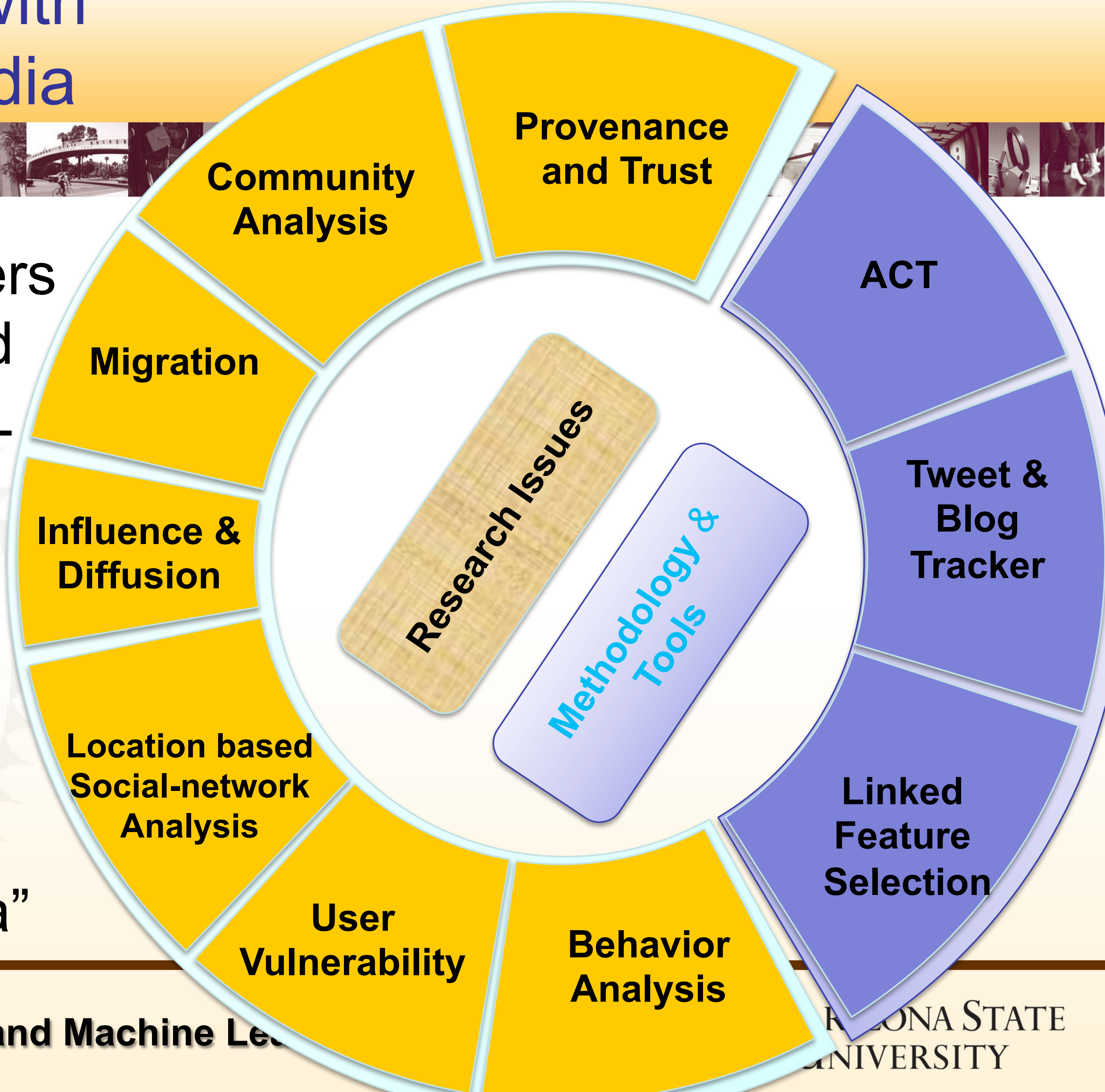
(a2) Q1s

(a1) Followers

(b1) Followers (b2) Q1s

Research with Social Media

- Recent papers can be found at our DMML Members' URLs
- A recent SI CFP on "Uncovering Deception in Social Media"



CAMBRIDGE
UNIVERSITY PRESS

FORTHCOMING FALL 2013!

Social Media Mining: An Introduction

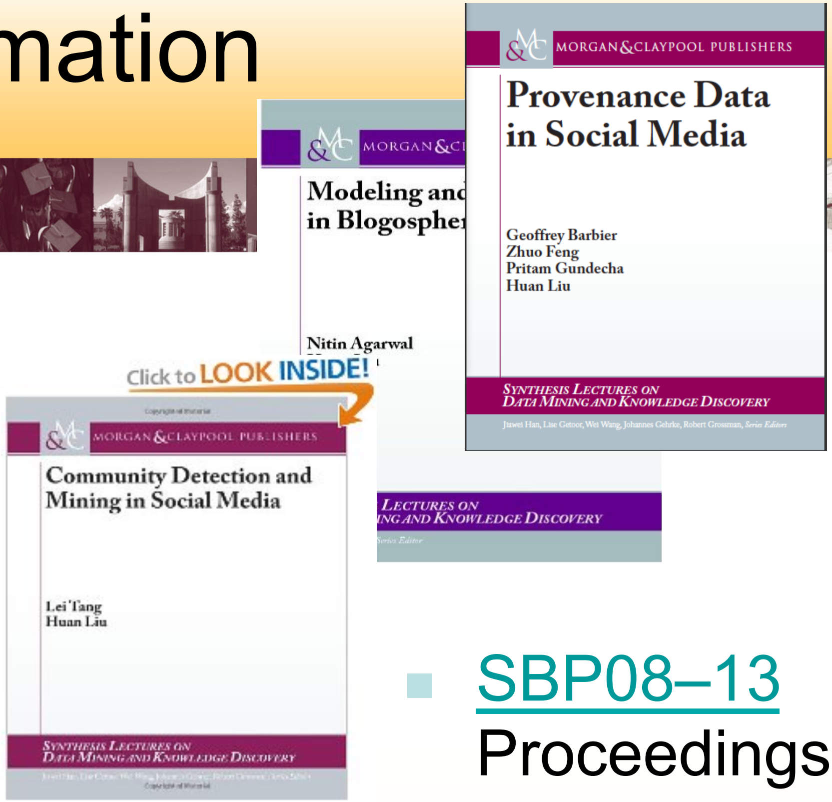
Huan Liu, Ali Abbasi, and Reza Zafarani, *Arizona State University*

This textbook goes from the basics to state-of-the-art, providing a single entry point for learning social media mining. It integrates the three key components of social media, social network analysis, and data mining, offering students and practitioners alike a comprehensive but focused understanding of this emerging multi-disciplinary field.

NOVEMBER 2013
HB ISBN: 9781107018853

www.cambridge.org/us

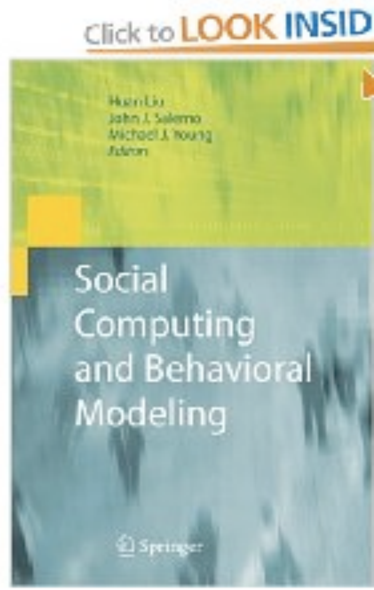
Additional Information



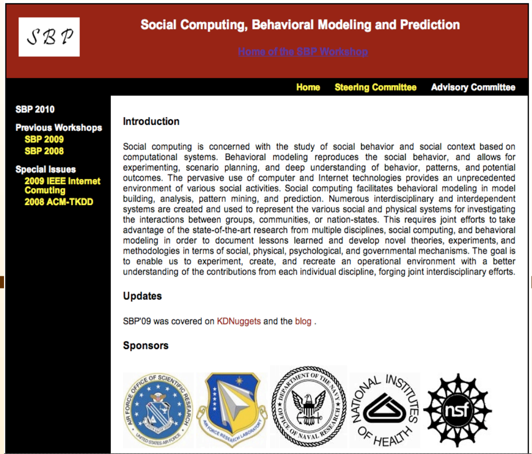
- Modeling and Data Mining in Blogosphere (2009)
- Community Detection and Mining (2010)
- Provenance Data in Social Media (2013)

- SBP08–13 Proceedings
- SBP14

April, D.C.



- SBP Conference Series



Thank You All

Acknowledgments: Projects are, in part, sponsored by ARO, NSF, and ONR; thanks to passionate and creative DMML members and our collaborators.

Thanks to ASONAM13 chairs for kind invitation and this wonderful opportunity

http://www.public.asu.edu/~huanliu