Data Mining &
Machine Learning

ARIZONA STATE UNIVERSITY

Some New Data Challenges for Data Science

Huan Liu

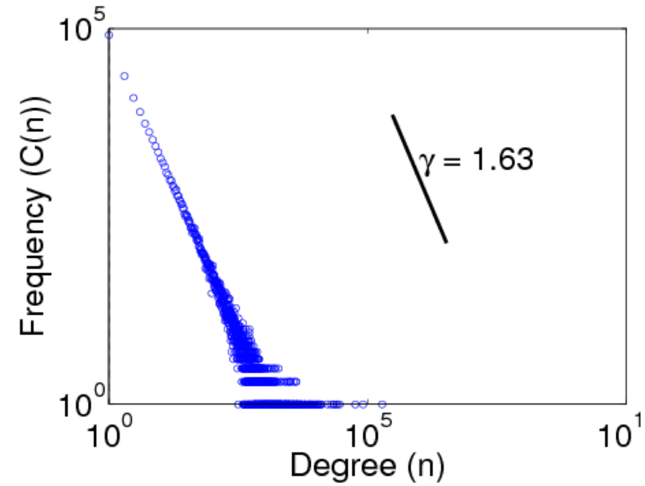
Ubiquitous Big Data and Data Science

- Abundant Data is Ubiquitous
  - It has changed the AI playing ground
- "Data is the New Oil"
  - AI finds a new lifeline from data
  - ***Data Science*** emerges from CS, Statistics, IS, etc.
- Recent success of AI is due to its use of ***data***
  - Machine Learning (e.g., Deep Learning)
- For any ML algorithm to work, data is key
  - We use social media data to illustrate Data Challenges

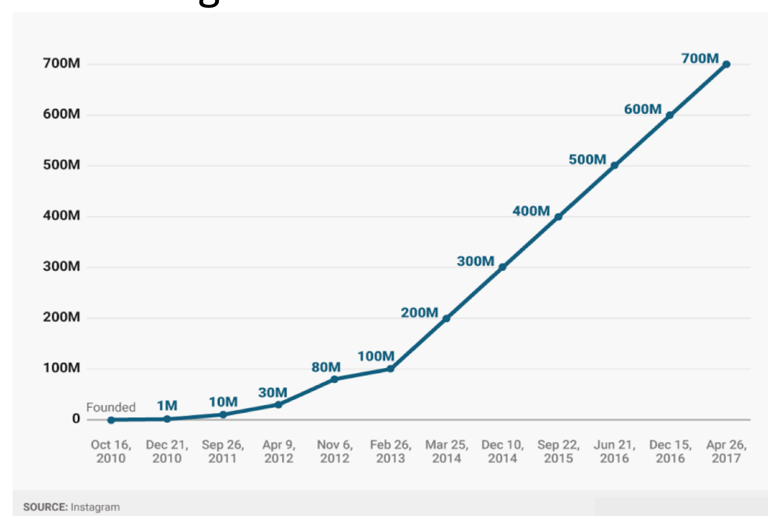
Social Media Data – A New Source of Big Data

- Twitter
  - 300 million users
  - 500 million tweets / day
  - 1% (5 million) released for research
- Facebook
  - 2 billion users
  - 422 million updates / day
  - 196 million photos / day
- Instagram
  - 700 million users
  - 80 million photos / day

Facebook Degree Distribution

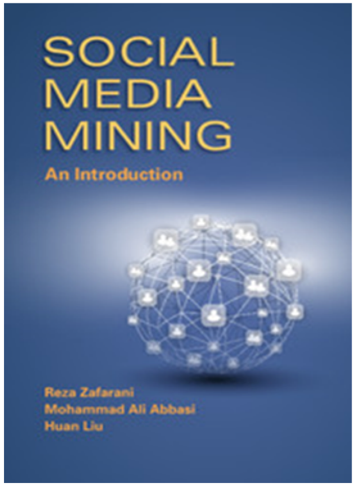


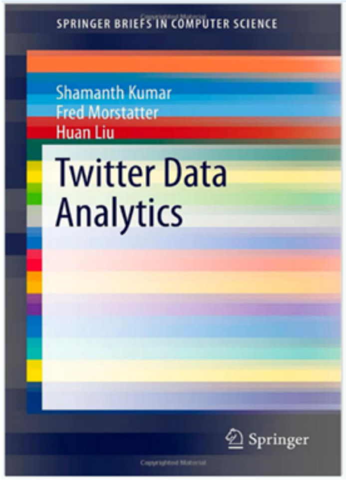
Instagram Users over Time



Mining Social Media Data

- Graph Theories
- Network Measures and Models
- Data Mining, NLP, and Visual Analytics
- Community Detection and Analysis
- Information Diffusion
- Influence and Homophily
- Recommender Systems
- Behavior Analytics
  - Sentiment Analysis





KDnuggets™ Top Stories, Jun 2, 2018

Featured Story

- **10 More Free Must-Read Books for Machine Learning and Data Science**

https://www.kdnuggets.com/2018/05/10-more-free-must-read-books-for-machine-learning-and-data-science.html

Social-Media-Data Challenges

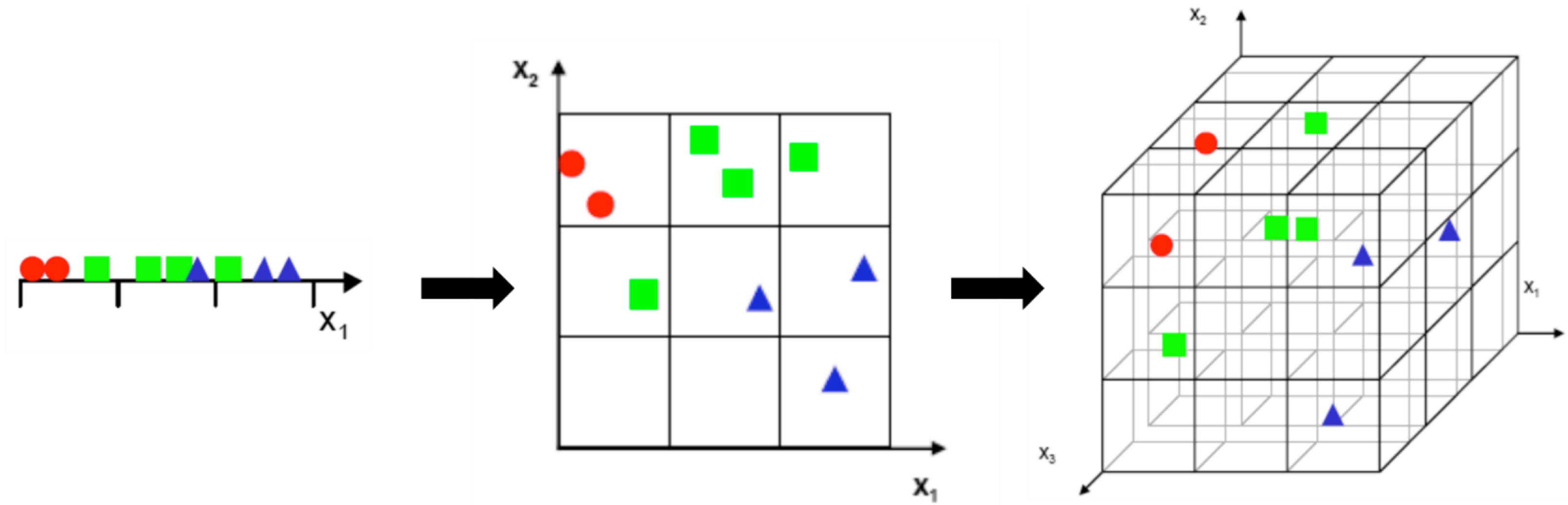
- SM data seems really big, is it really so?
  - How can we make data *bigger*?
- Data can be revealing, where is our privacy?
  - Do we have to make a trade-off between *privacy* and *utility*?
- An ultimate challenge for our research to be accepted or reproducible is …?
  - How can we *evaluate without ground truth*?

Making Big Data "Bigger"

- What is big data?
  - A conventional answer is 4Vs
  - A practitioner's answer is more nuanced
- Big data can be actually *little* or *thin*
- When small data alone is insufficient, ***we need to find more or bigger data***
  - Make little data bigger
  - Make thin data thicker

Curse of Dimensionality: Required Samples

- Sparsity becomes exponentially worse as dimensionality increases
  - Conventional distance metric becomes ineffective as far and near neighbors have similar distances

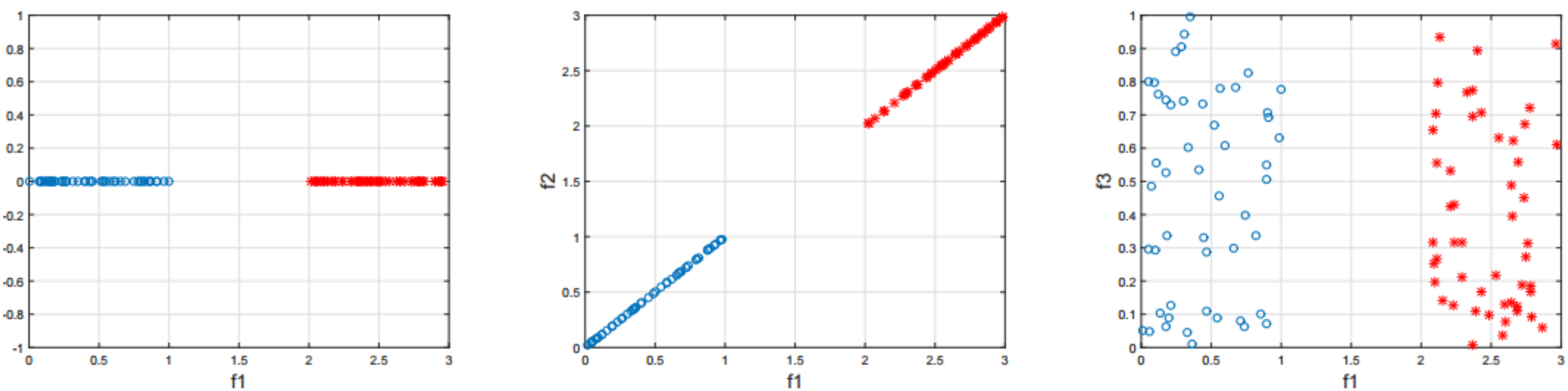


3 samples per unit region **1 sample per region** **1/3 sample per region**

http://nikhilbuduma.com/2015/03/10/the-curse-of-dimensionality/

Relevant, Redundant and Irrelevant Features

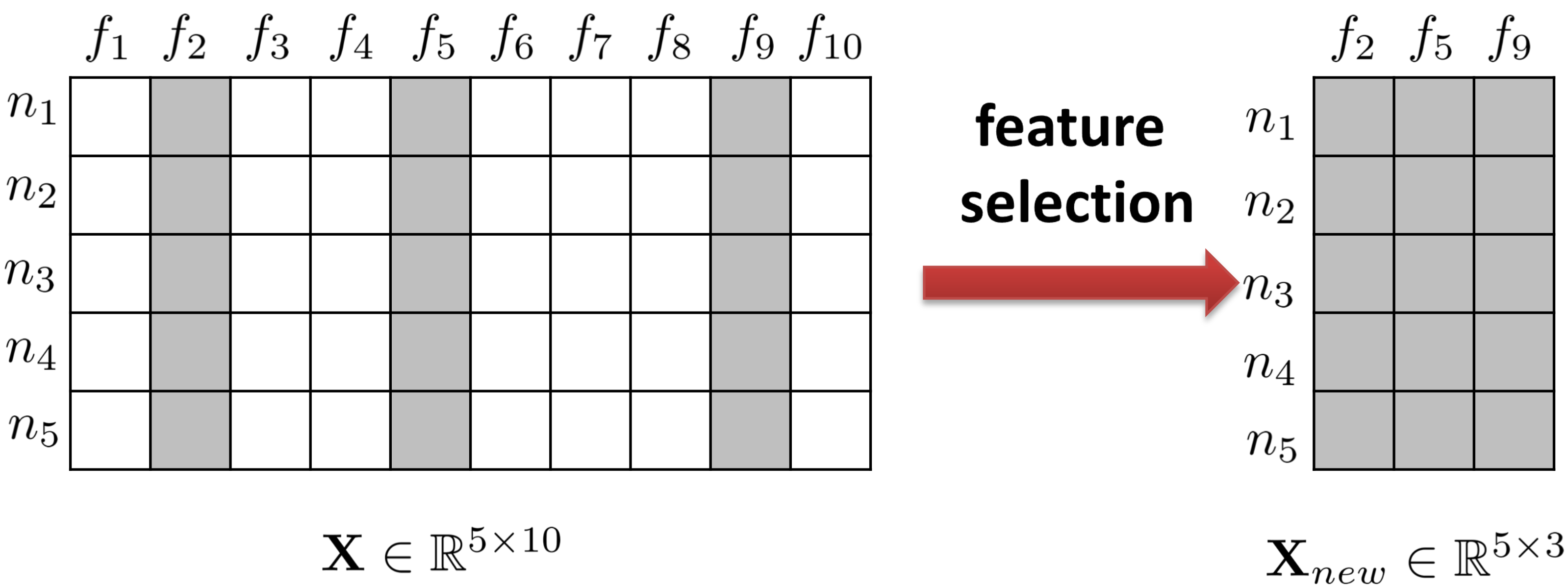
- Feature selection retains relevant features for learning and removes redundant or irrelevant ones
- For a binary classification task below, f_1 is relevant, f_2 is redundant given f_1, and f_3 is irrelevant



(a) relevant feature f_1 (b) redundant feature f_2 (c) irrelevant feature f_3

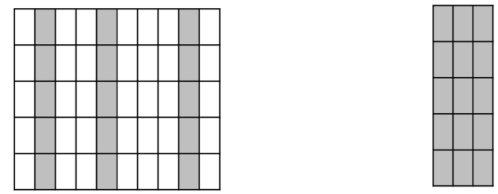
Feature Selection Can Make Data Bigger

Feature selection finds an 'optimal' subset of relevant features from the original high-dimensional data given a certain criterion

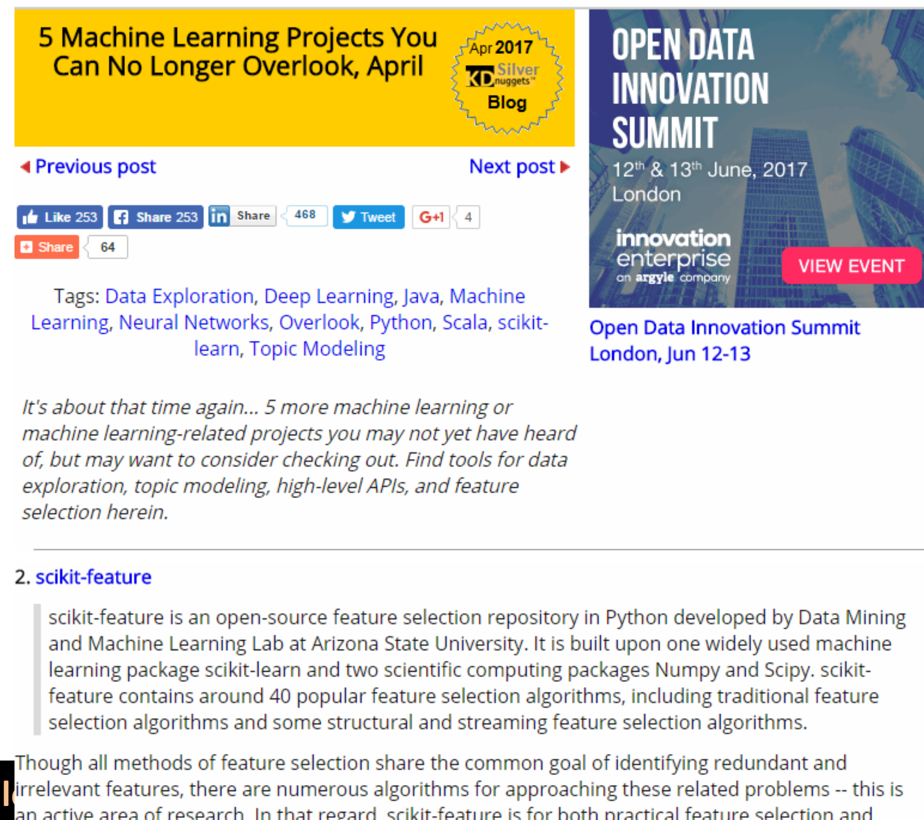


Feature Selection and scikit-feature

- Feature selection can make data `bigger’
  - Assuming all binary attribute values in our toy example
  - Before FS, $5/2^{10} = 5/1024$, after FS, $5/2^3 = 5/8$
- Does FS always work?
  - Yes, for most high-d data
- Where can we find it?
- ***scikit-feature***, an open-source repository in Python



$\mathbf{X} \in \mathbb{R}^{5\times 10}$ $\mathbf{X}_{new} \in \mathbb{R}^{5\times 3}$


5 Machine Learning Projects You Can No Longer Overlook, April

Apr 2017 KDnuggets Silver Blog

Previous post Next post

Tags: Data Exploration, Deep Learning, Java, Machine Learning, Neural Networks, Overlook, Python, Scala, scikit-learn, Topic Modeling

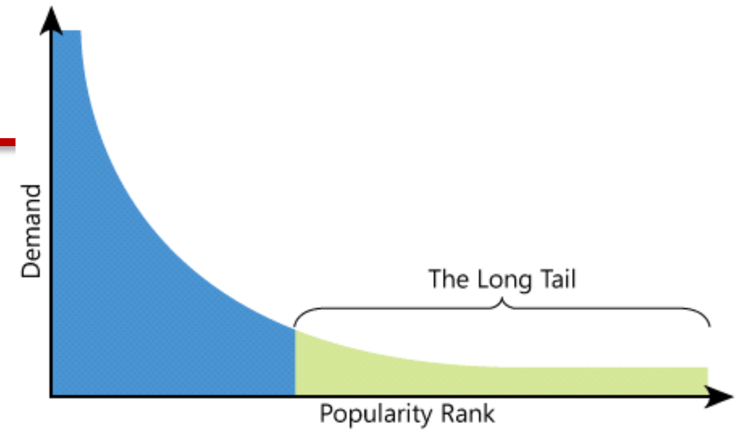
It's about that time again... 5 more machine learning or machine learning-related projects you may not yet have heard of, but may want to consider checking out. Find tools for data exploration, topic modeling, high-level APIs, and feature selection herein.

2. scikit-feature

> scikit-feature is an open-source feature selection repository in Python developed by Data Mining and Machine Learning Lab at Arizona State University. It is built upon one widely used machine learning package scikit-learn and two scientific computing packages Numpy and Scipy. scikit-feature contains around 40 popular feature selection algorithms, including traditional feature selection algorithms and some structural and streaming feature selection algorithms.

Though all methods of feature selection share the common goal of identifying redundant and irrelevant features, there are numerous algorithms for approaching these related problems -- this is an active area of research. In that regard, scikit-feature is for both practical feature selection and

Making Thin Data Thicker



- Most people like many of us are in the long tail
  - Our data is thin or sparse
  - With little data, machine learning is powerless
- Social media data offers new opportunities
  - Multiple facets: posts, profile, linked information
  - Multiple platforms that offer different functions
- Two illustrative cases
  - Selecting features using ***social network*** information
  - Connecting users ***across*** social media sites

Online User Data: Utility vs. Privacy

- Users conduct numerous online activities
- Each user is leaving their data traces
- Their data helps improve personalized services

Browsing Histories Can also Reveal Privacy

- Adversaries can infer different types of personally identifiable information
- Web browsing history data is finger-printable
  - New attacks that map a given history to a social media profile
- Users can become vulnerable to various harms

6

The Relationship btw Privacy and Utility

- Conventional solutions often make a trade-off between privacy and utility
- Reduced utility can result in decreased quality of online personalized services
- Hence, the dilemma of privacy and utility
  - Can we have both?

Attacks via Web Browsing History

Threat Model: Given *u*'s browsing history $\mathcal{H}^u = \{l_1, \ldots, l_n\}$, map *u* to a social media profile based on the links in its feed

Twitter feed

Browsing history

https://facebook.com

http://cs246.stanford.edu

http://voxeu.org/article/...

Challenges in Anonymizing Browsing Histories

- How privacy and utility should be defined in this context?
- How many links should be added?
- What links should be added?

Measuring Privacy

- The more ambiguous a user's interests are, the harder it is for the adversary to infer her characteristics
- Entropy is used as a measure of ambiguity

$$Privacy(p_u) = -\sum_{j=1}^{m} p_{uj} \log p_{uj}$$

The higher the entropy, the higher the privacy

Topic probability distribution

Measuring Utility Loss

- The more difference between user's topic distribution before and after anonymization, the more lost utility of her browsing history

$$utility\_loss(p_u, \hat{p_u}) = 0.5 \times (1 - sim(p_u, \hat{p}_u))$$

Topic probability distribution after anonymization

$$sim(p_u, \hat{p_u}) = \frac{p_u \cdot \hat{p}_u}{||p_u|| \cdot ||\hat{p}_u||}$$

PBooster Algorithm

1. Topic Selection

– Select a subset of topics and calculate the number of links that should be added to each topic

$$a^* = \mathrm{argmax}_a \, G(p_u, \hat{p_u}, \lambda)$$

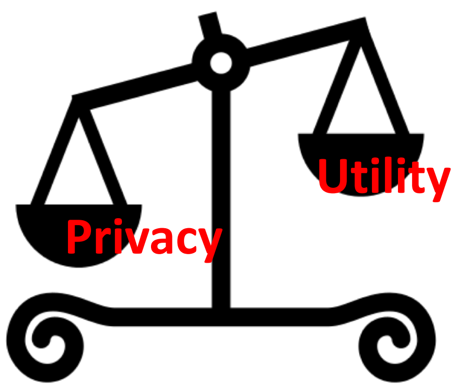
Beauty	Sport	Food	Politics
5	0	3	10

2. Link Selection

– Select a proper set of links that corresponds to the identified topics and their numbers

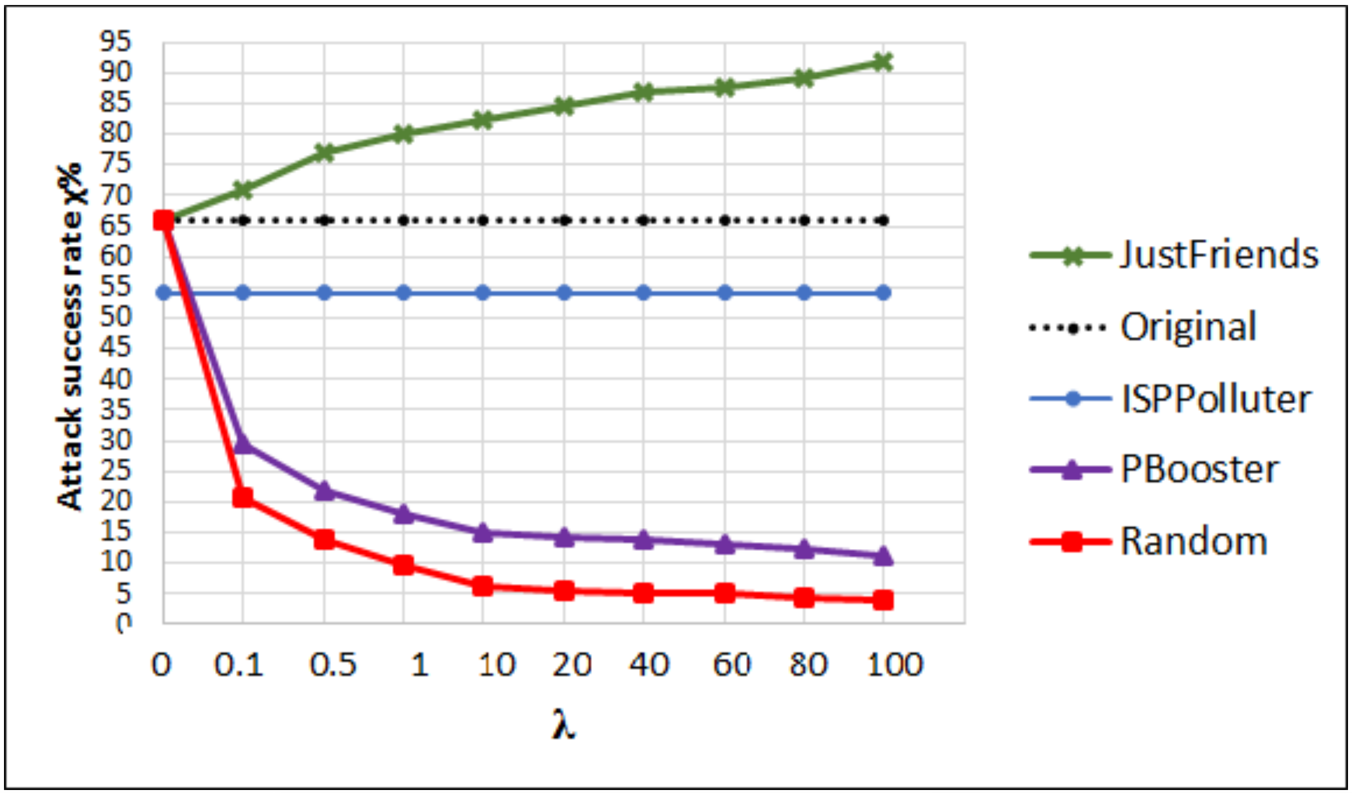
Experimental Evaluation

- To answer the following questions:
  1. Can PBooster help protect user privacy?
  2. How does PBooster change the utility, or the quality of online services?
  3. Do we have to make a trade-off between privacy and utility?
     - Does PBooster make a difference?



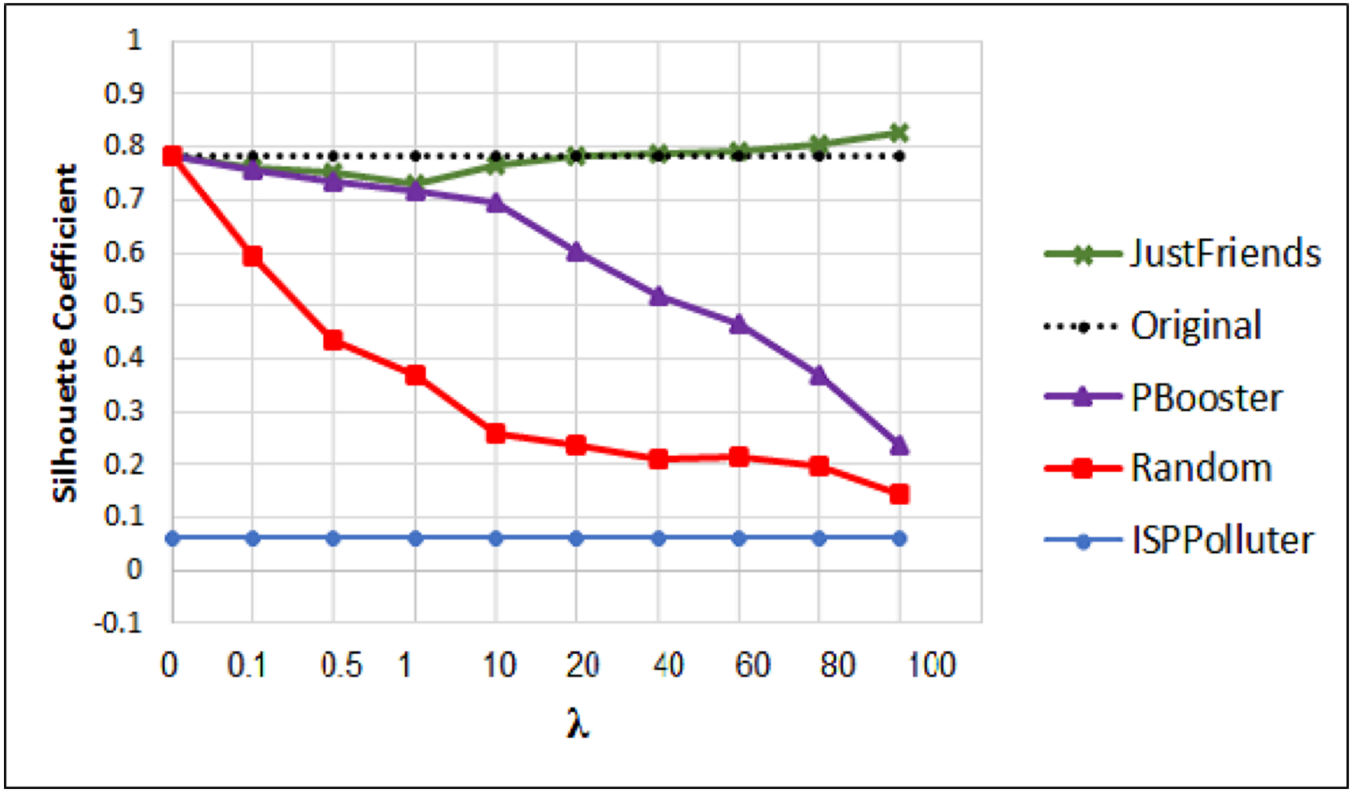
Privacy Analysis

- **Privacy evaluation:** Deploy de-anonymization attack
- **Evaluation metric**: Attack success rate = $\frac{n_c}{N}$
  - Attack is successful if the user is among the top 10 results

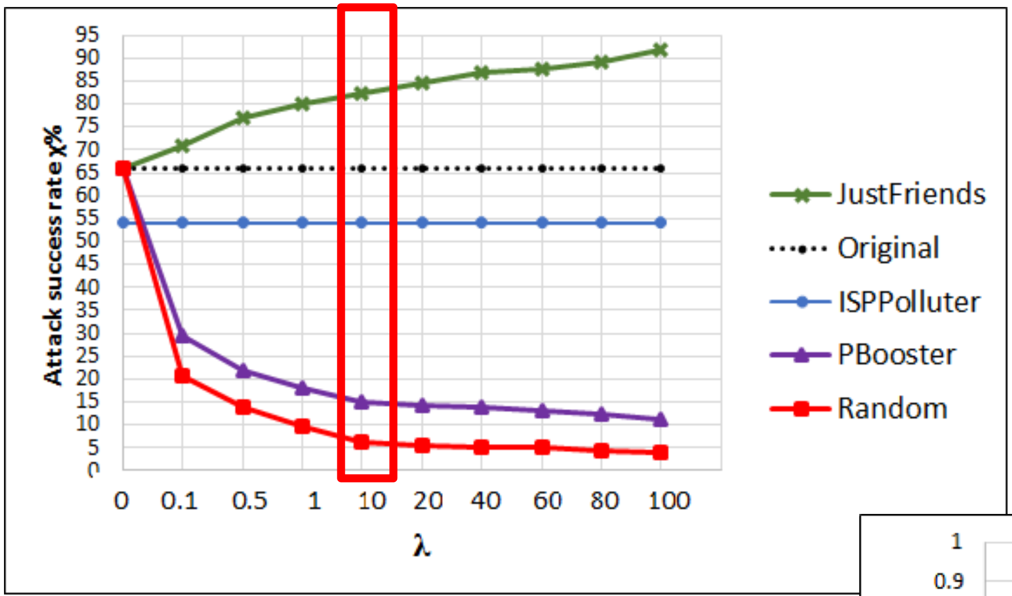


Utility Analysis

- **Utility evaluation:** Cluster users with k-means based on topic probability distributions into k = 5 groups
- **Evaluation metric**: Evaluate quality of generated clusters with Silhouette Coefficient ranges from [-1,1]



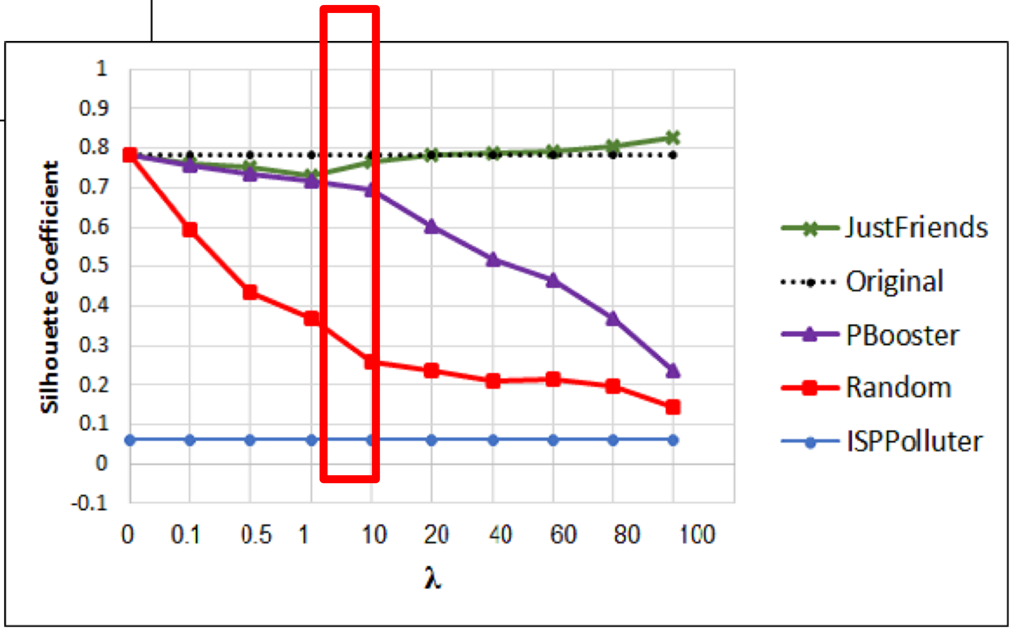
Sweet Spots for High Privacy and Utility



Privacy Evaluation: Deploy an existing de-anonymization attack

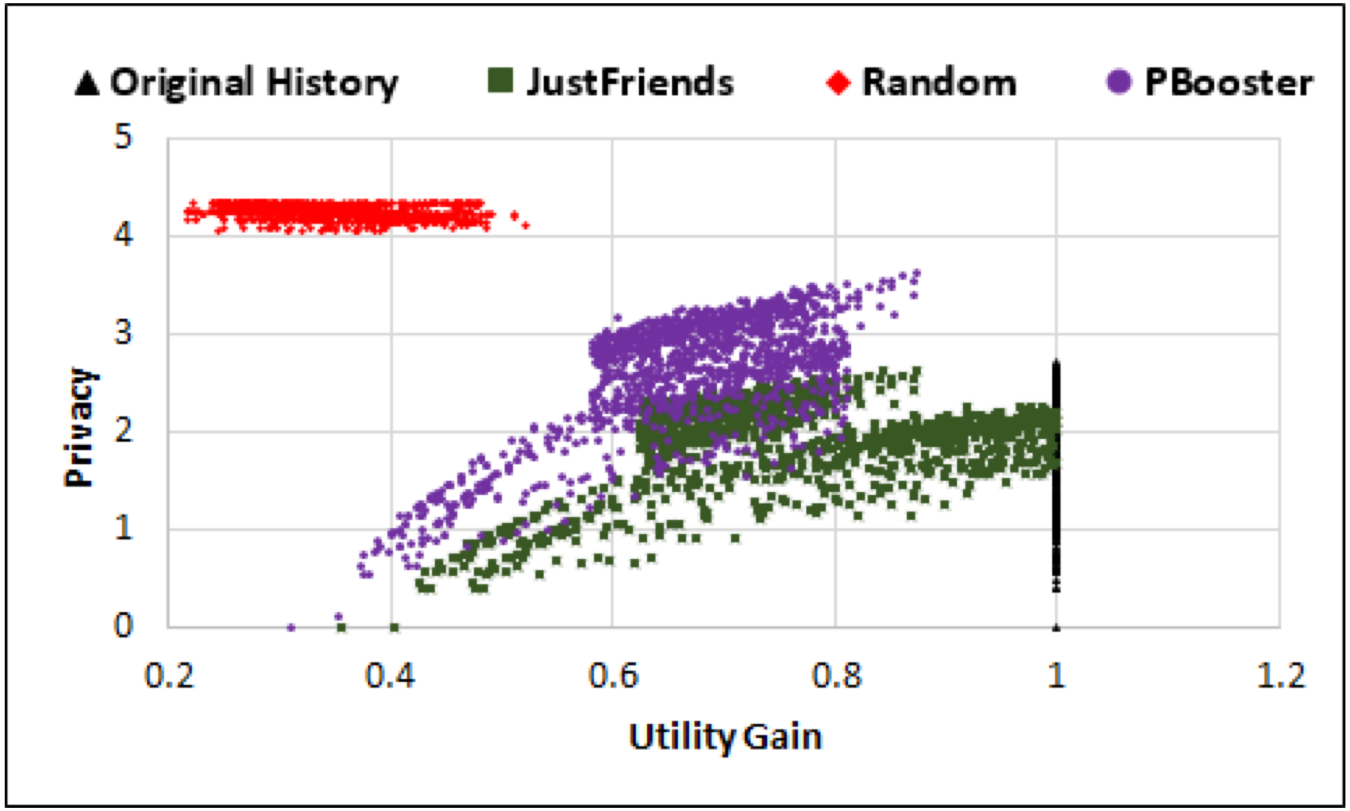
Attack is successful if the user is among the top 10 results

Utility Evaluation: Cluster users with k-means based on topic probability distributions

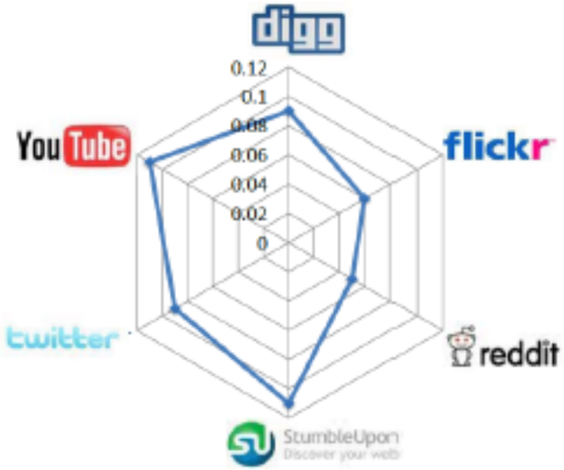


Privacy-Utility Trade-off: Is it Necessary?

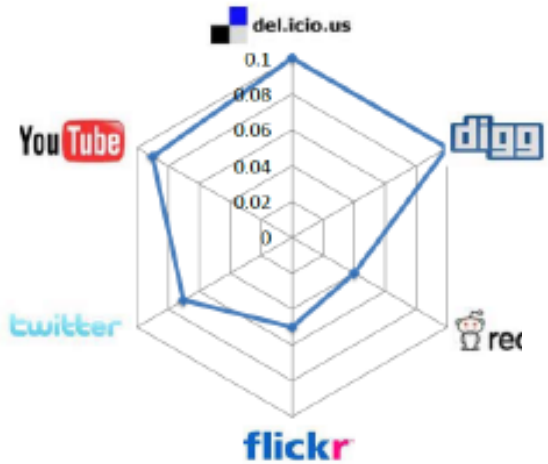
Plotting privacy and utility gain values for each user after applying different approaches over histories



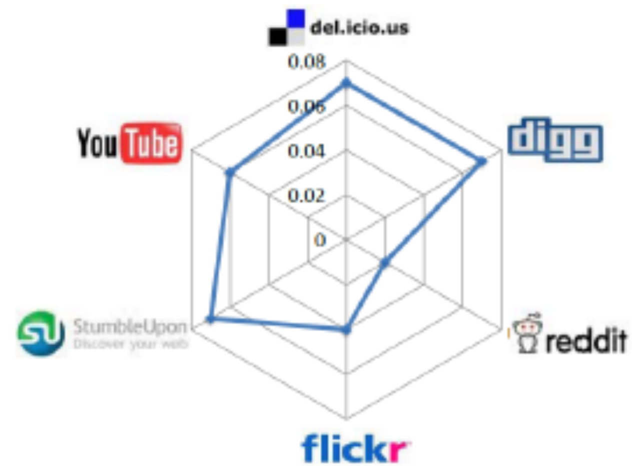
No Ground Truth: Migration on Social Media Platforms



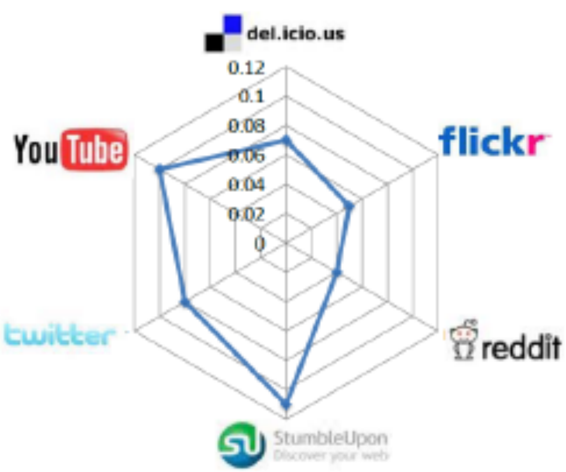
(a) Delicious



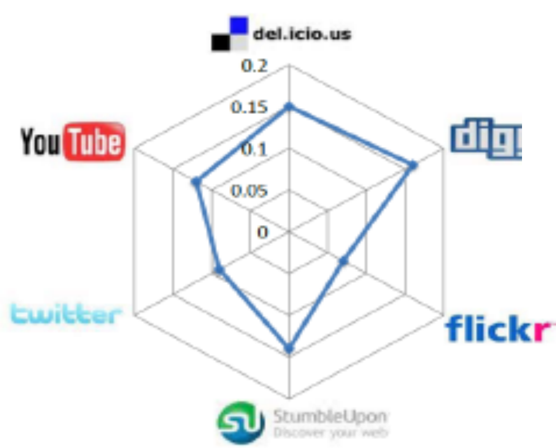
(e) StumbleUpon



(f) Twitter

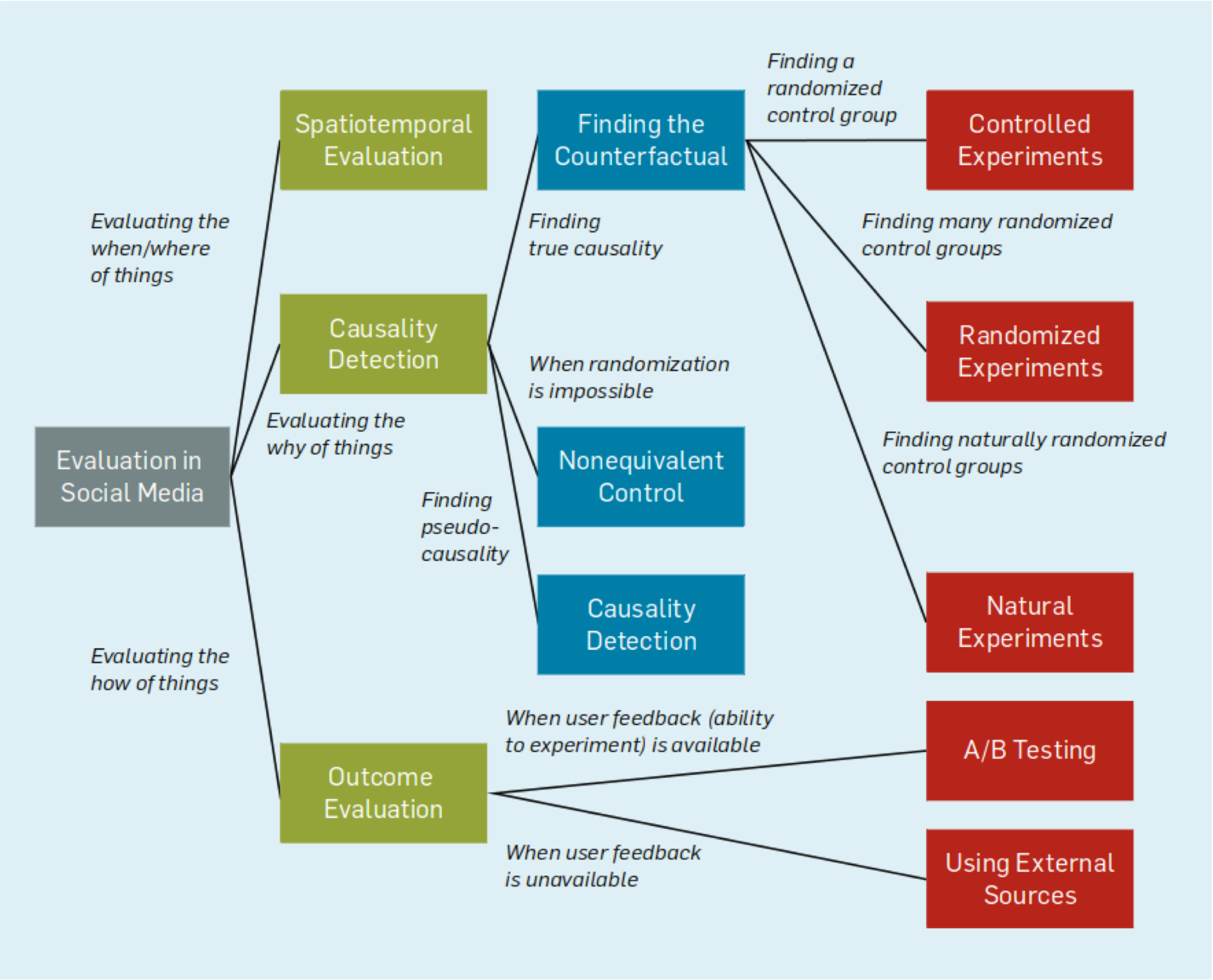


(b) Digg



(d) Reddit

Evaluation without Ground Truth

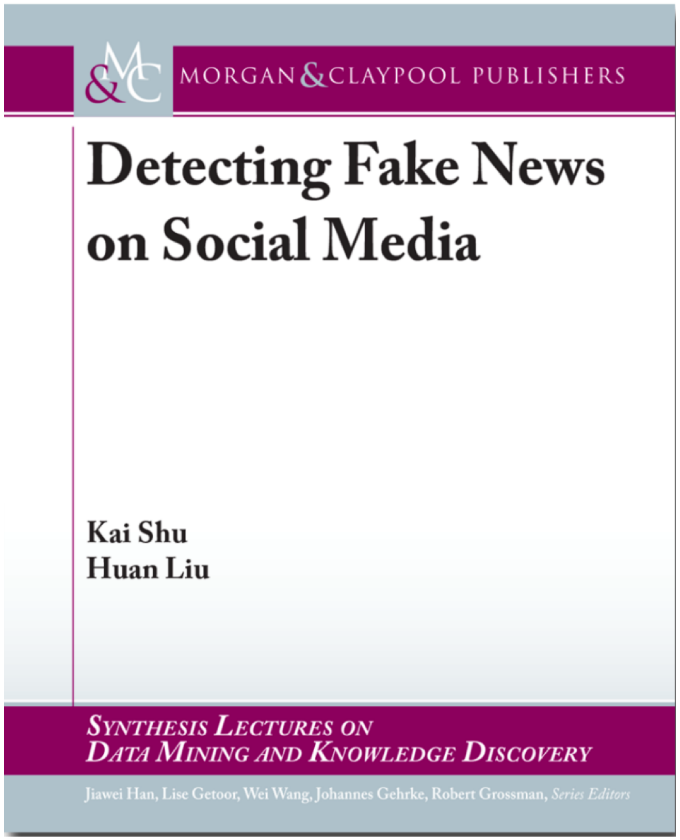


The CACM article can be found at dl.acm.org

Social Media Data Challenges

- SM data seems really big, is it really so?
  - How can we make data *bigger*?
- Data can be revealing, where is our privacy?
  - Do we have to make a trade-off between *privacy* and *utility*?
- An ultimate challenge for our research to be accepted or reproducible is …?
  - How can we *evaluate without ground truth*?

Call for Authors



CALL FOR AUTHORS

Understanding Artificial Intelligence (AI): Natural Language Processing (NLP), Causality, Fairness, and Ethics series

A NEW BOOK SERIES FROM CRC PRESS

Series Editor: Huan Liu

This series aims to provide a conducive platform for researchers and educators to share their findings and lessons learned, focusing on the understanding of AI from both the research perspective and the users' perspective. On the former, we advocate the understanding of AI by advancing techniques and theories in Natural Language Processing (NLP), ChatBots, Bias in Data, Knowledge, Representation, and Evaluation, Fairness, Transparency, Deception, and Explainable AI; on the latter, we endeavor to demystify what AI can do, what responsible AI is via discussion on Bias, Ethics, and Education.

This series will include books on topics that are of interest to researchers and users of AI. The opacity of AI techniques raises unnecessary worries and stimulates wild speculation of AI's imagined un-reined power. This series gathers state-of-the-art research on understanding AI to help demystify AI techniques and algorithms via research and education on NLP, causality, fairness, and ethics.

Consideration will be given to a broad range of textbooks, monographs, reference works, and handbooks that appeal to academics, practitioners, and students. We encourage the submission of all relevant topics in the areas of: AI Ethics, AI and Fairness, Human and Machine Teaming, Machine Learning, Privacy in AI, AI in Healthcare, AI in Personalized Medicine, Knowledge and Data in AI, Natural Language Processing, Causal Learning, to name just a few topics of interest.

PROPOSALS MAY BE SUBMITTED TO:

Huan Liu
Professor of Computer Science and Engineering
Arizona State University
huanliu@asu.edu
http://www.public.asu.edu/~huanliu

Cindy Renee Carelli
Executive Editor
CRC Press – Taylor & Francis Group
cindy.carelli@taylorandfrancis.com

www.crcpress.com | email: orders@crcpress.com

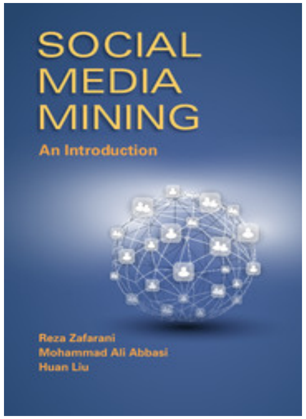
CRC Press
Taylor & Francis Group

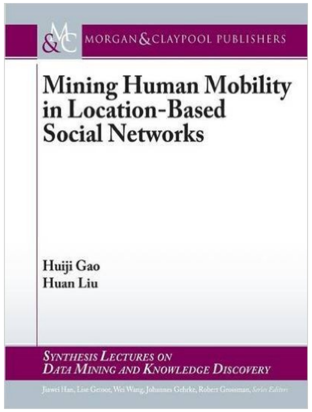
Use Code **FAKENEWS** during checkout, 30% Discount

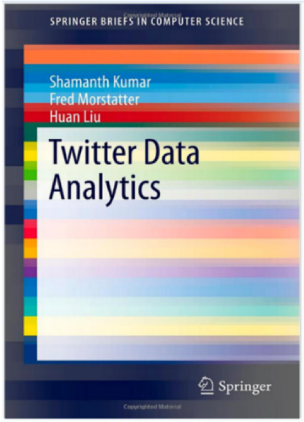
www.morganclaypoolpublishers.com/fakenews

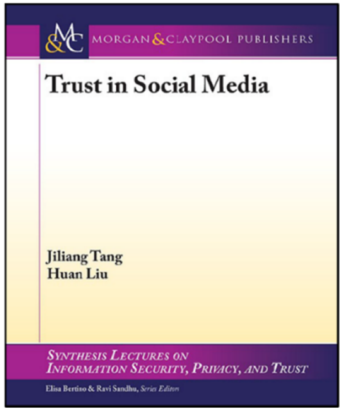
THANKS with Repositories, Surveys, and Books

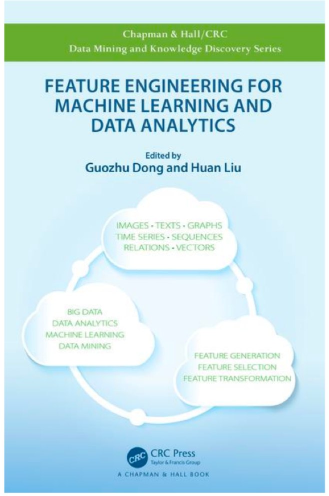
- ***scikit-feature*** – an open source feature selection repository in Python
- Social Computing Repository
- Two Recent Surveys
  - **Learning Causality with Data: Problems & Methods**
  - **Privacy in Social Media: Identification, Mitigation, …**











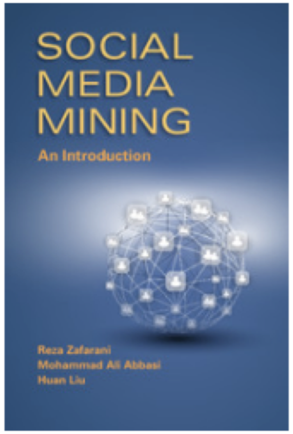
http://www.public.asu.edu/~huanliu

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/

Challenges with Social Media Data

- Social media data seems really big, but why are we often still short of data?
  - How can we make data ‘*bigger*’?
- Data is power, so it can produce any result
  - Can we *algorithmically* evaluate the results from big data?
- We don’t know what we don’t know
  - How can we know if our result of social media analysis is of any value?

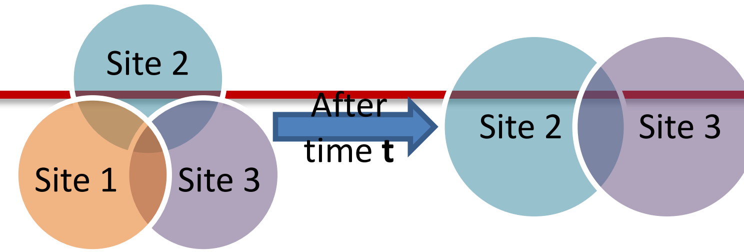
Addressing Don't-Know-Don't-Know Problems

- When collecting data, we often ***don't know*** when we have a sufficient amount
  - We don't know *when to stop* collecting, though we can't collect forever
- A dilemma in studying ***migration*** on social media :
  - If we know its existence, no need for the study
  - If we ***don't know***, how can we verify the result?

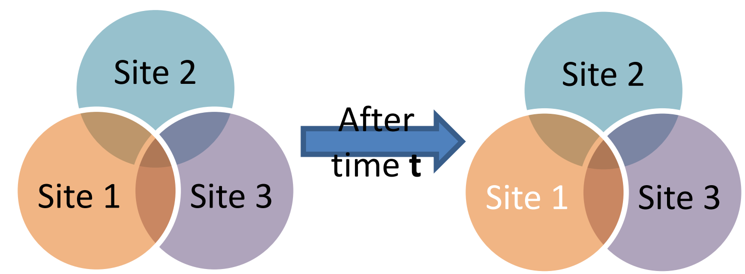
Illustrative Examples of DNDN

1. When-to-Stop Dilemma: Collecting data forever vs. having credible patterns
   – How much data vs. how credible
2. Is There Migration on Social Media?
   – Users are a primary source of revenue
      • Ads, Recommendations, Brand loyalty
   – New SM sites need to *attract* users for expansion
   – Existing SM sites need to *retain* their users
   – ***Competiting for attention*** entails the discovery of migration patterns

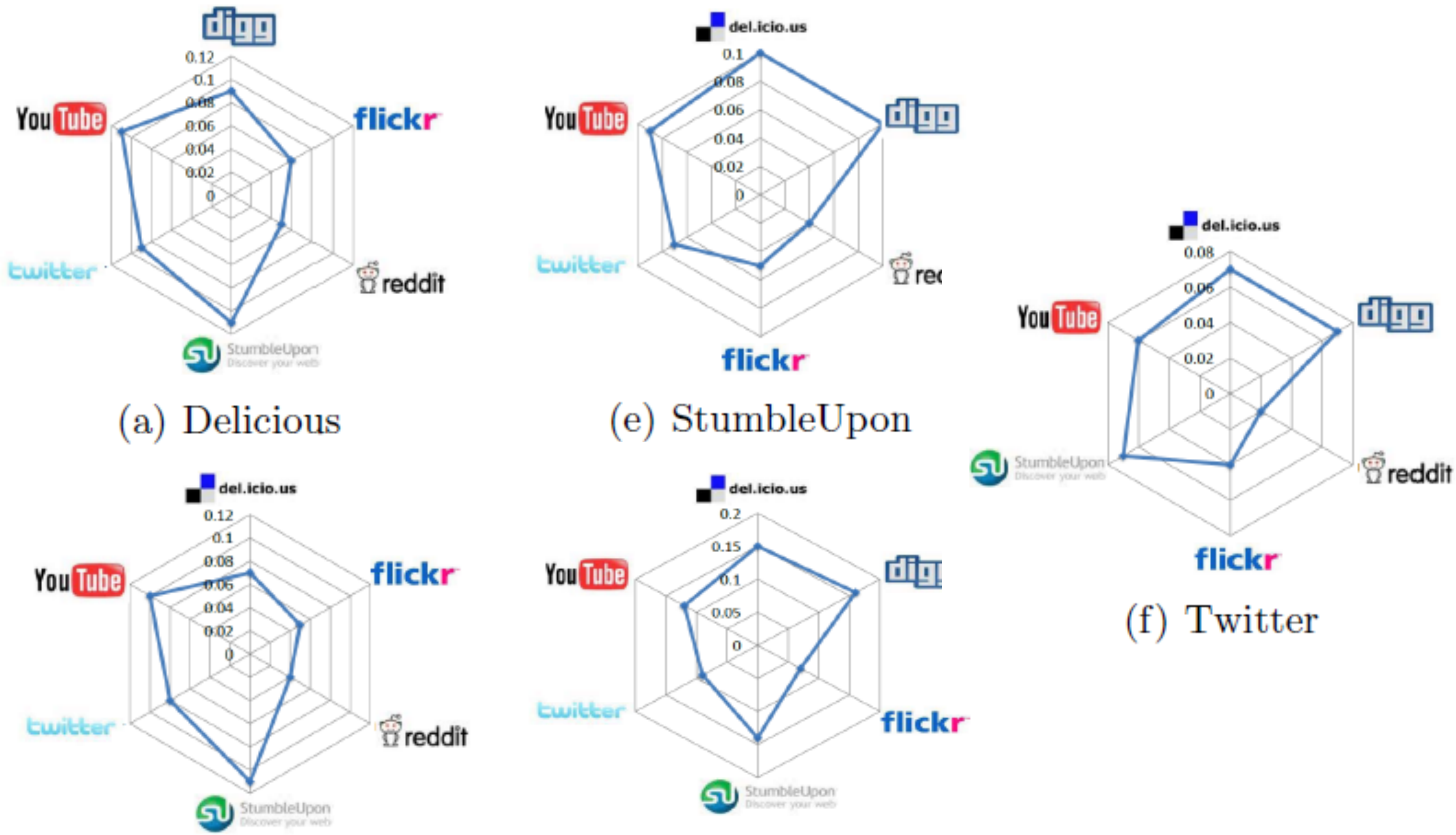
Migration on Social Media



- Site Migration
  - Users leave a site by profile deletion or profile removal
  - Difficult to convince a user who left to return
  - Hard to study these users cross sites because we need their registration information
- **Attention Migration**
  - Users become inactive on a site
  - A harbinger for site migration
  - Can be detected by observing *user activities* across sites
  - Can take action to prevent site migration after understanding migration patterns



Patterns from Observation



(a) Delicious

(e) StumbleUpon

(f) Twitter

(b) Digg

(d) Reddit

Do We Know What We Didn't Know?

- If a pattern is significant, it is valid
  - Significant differences observed in StumbleUpon, Twitter, and YouTube
- When to stop?

  Stop when we are certain, continue otherwise

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

Revisit Challenges in Acquiring SM Intelligence

- Social media data is obviously big, but why are we often still short of data?
  - How can we make data `*bigger*'?
- Data is power, so it can produce any result
  - Can we *algorithmically* evaluate the results from big data?
- We don't know what we don't know
  - How can we know if our result of social media analysis is of any value?