Social Media Mining

Behavior Analytics

Dear instructors/users of these slides:

Please feel free to include these slides in your own material, or modify them as you see fit. If you decide to incorporate these slides into your presentations, please include the following note:

R. Zafarani, M. A. Abbasi, and H. Liu, *Social Media Mining: An Introduction*, Cambridge University Press, 2014.
Free book and slides at **http://socialmediamining.info/**

or include a link to the website:
http://socialmediamining.info/

Examples of Behavior Analytics

- What motivates users to join an online group?

- When users abandon social media sites, where do they migrate to?

- Can we predict box office revenues for movies from tweets?

Behavior Analysis

- To answers these questions we need to **analyze** or **predict** behaviors on social media.

- Users exhibit different behaviors on social media:
  - As individuals, or
  - As part of a broader collective behavior.

- When discussing individual behavior,
  - Our focus is on one individual.

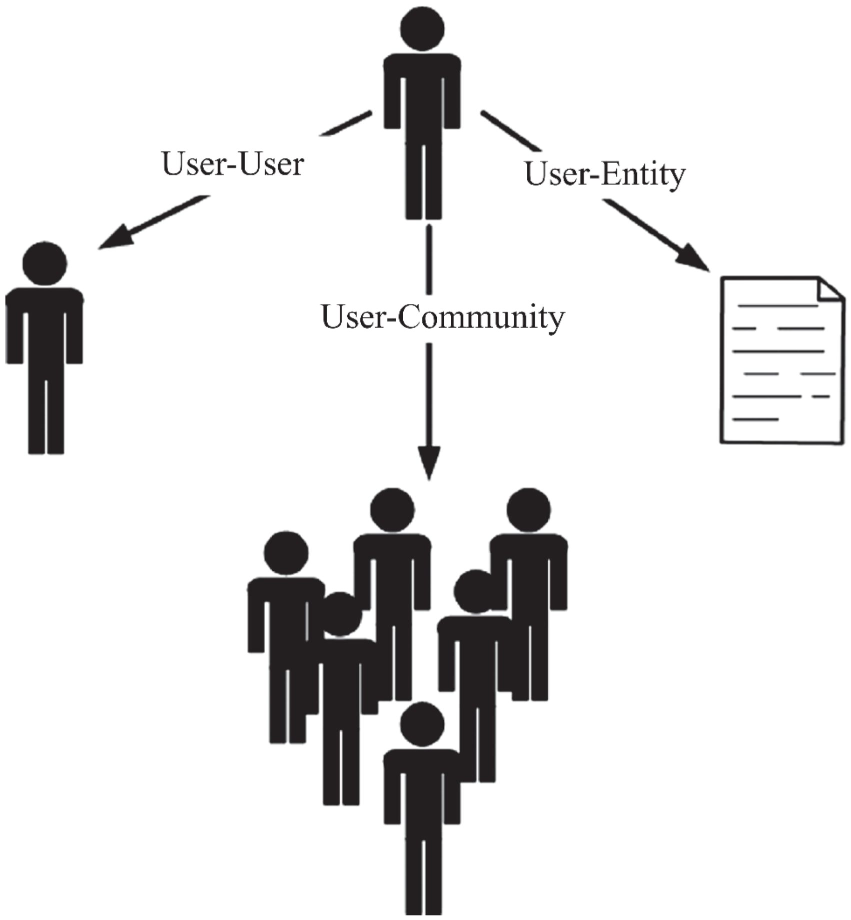
- Collective behavior emerges when ***a population of individuals behave in a similar way with or without coordination or planning***.

Our Goal

To analyze, model, and predict individual and collective behavior

Individual Behavior

Types of Individual Behavior



- **User-User**
  (link generation)
  – befriending, sending a message, playing games, following, or inviting
- **User-Community**
  – joining or leaving a community, participating in community discussions
- **User-Entity**
  (content generation)
  – writing a post
  – posting a photo

I. Individual Behavior Analysis

Example: Community Membership in Social Media

- Why do users join communities?
  - Communities can be implicit:
    - Individuals buying a product as a community, and
    - People buying the product for the first time as individuals joining the community.
  - **What factors affect the community-joining behavior of individuals?**

- We can observe users who join communities
  - **Determine factors that are common among them**

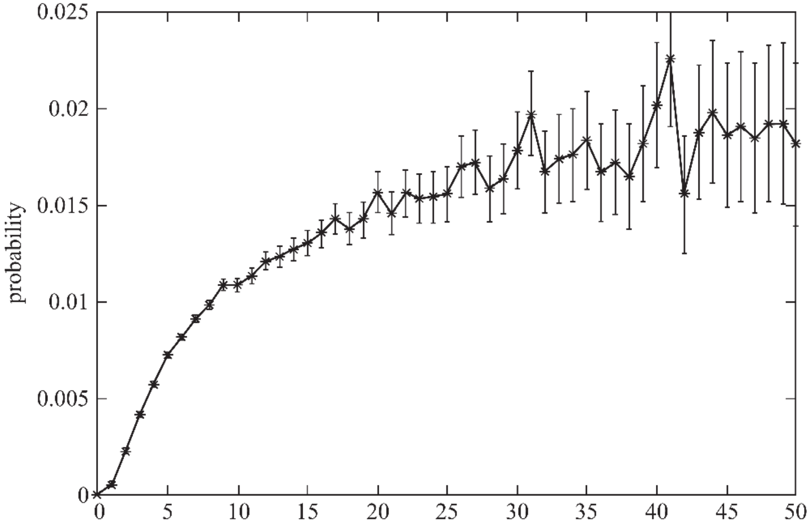
- To observe users, we require
  - A population of users,
  - A community C, and
  - Community membership info (users who are members of C)

- To distinguish between users who have already joined the community and those who are now joining it,
  - We need community memberships at two times t_1 and t_2, with $t_2 > t_1$
  - At t_2, we find users who are members of the community, but were not members at t_1
    - These new users form the subpopulation that is analyzed for community-joining behavior.

Community Membership in Social Media

Hypothesis:

  - individuals are inclined toward an activity when their friends are engaged in the same activity.



- A factor that plays a role in users joining a community is the number of their friends who are already members of the community.

- In data mining terms,
  - number of friends of an individual in a community
  - A **feature** to predict whether the individual joins the community (i.e., **class attribute**).

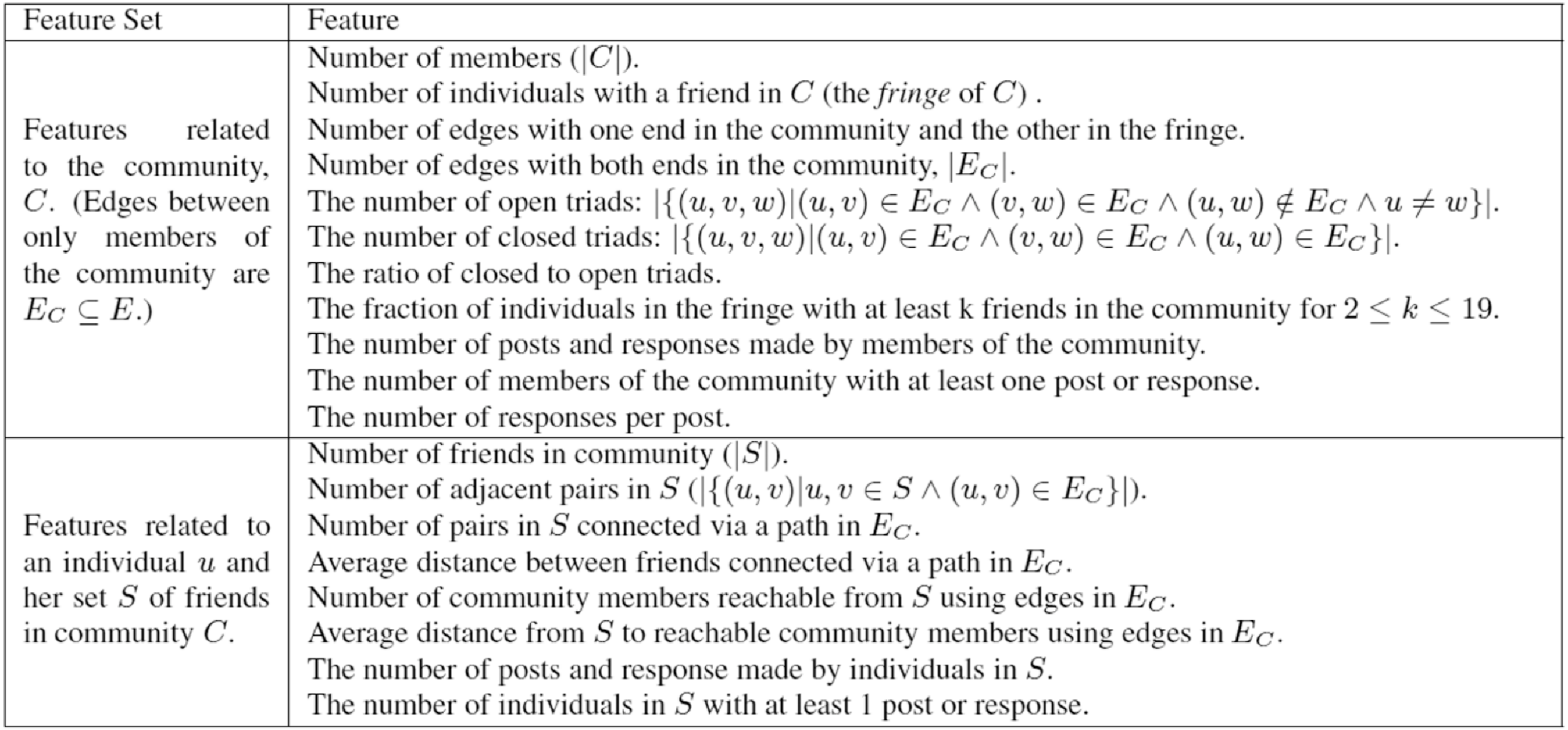
Number of Friends

vs

Probability of Joining a Community

Backstrom, L., Huttenlocher, D., Kleinberg, J., & Lan, X. (2006, August). Group formation in large social networks: membership, growth, and evolution. In Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 44-54). ACM.

Even More Features



Feature Set	Feature
Features related to the community, C. (Edges between only members of the community are $E_C \subseteq E$.)	Number of members ($\lvert C\rvert$). Number of individuals with a friend in C (the *fringe* of C) . Number of edges with one end in the community and the other in the fringe. Number of edges with both ends in the community, $\lvert E_C\rvert$. The number of open triads: $\lvert\{(u,v,w)\lvert(u,v) \in E_C \wedge (v,w) \in E_C \wedge (u,w) \notin E_C \wedge u \neq w\}\rvert$. The number of closed triads: $\lvert\{(u,v,w)\lvert(u,v) \in E_C \wedge (v,w) \in E_C \wedge (u,w) \in E_C\}\rvert$. The ratio of closed to open triads. The fraction of individuals in the fringe with at least k friends in the community for $2 \leq k \leq 19$. The number of posts and responses made by members of the community. The number of members of the community with at least one post or response. The number of responses per post.
Features related to an individual u and her set S of friends in community C.	Number of friends in community ($\lvert S\rvert$). Number of adjacent pairs in S ($\lvert\{(u,v)\lvert u,v \in S \wedge (u,v) \in E_C\}\rvert$). Number of pairs in S connected via a path in E_C. Average distance between friends connected via a path in E_C. Number of community members reachable from S using edges in E_C. Average distance from S to reachable community members using edges in E_C. The number of posts and response made by individuals in S. The number of individuals in S with at least 1 post or response.

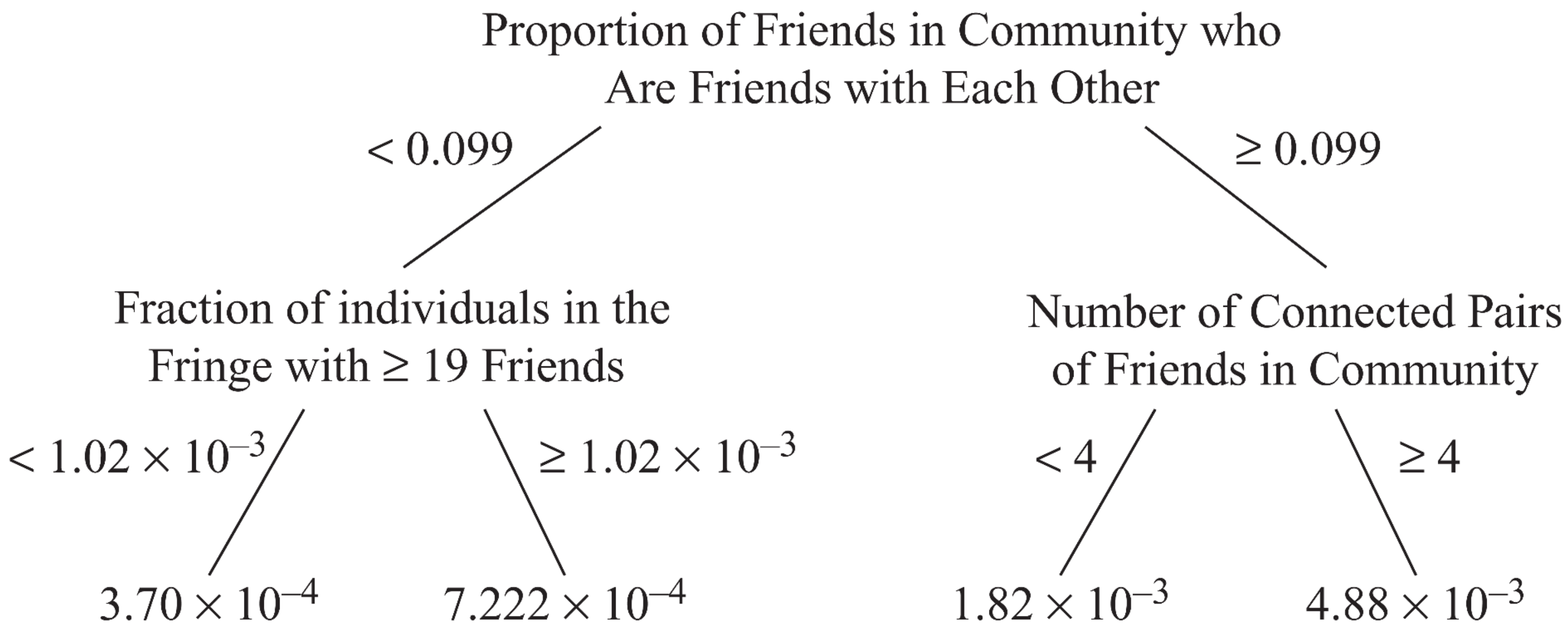
Feature Importance Analysis

Which feature can help best determine whether individuals will join or not?

I. We can use any feature selection algorithm, *or*

II. We can use a classification algorithm, such as decision tree learning
   - Most important Features are **ranked higher**

Decision Tree for Joining a Community


Proportion of Friends in Community who Are Friends with Each Other
< 0.099
≥ 0.099
Fraction of individuals in the Fringe with ≥ 19 Friends
Number of Connected Pairs of Friends in Community
$< 1.02 \times 10^{-3}$
$\geq 1.02 \times 10^{-3}$
< 4
≥ 4
3.70×10^{-4}
7.222×10^{-4}
1.82×10^{-3}
4.88×10^{-3}

Are these features well-designed?

- We can evaluate using classification performance metrics

Behavior Analysis Methodology

- **An observable behavior**
  - The behavior needs to be observable
  - E.g., accurately observing the joining of individuals (and possibly their joining times)

- **Features:**
  - Finding data features (covariates) that may or may not affect (or be affected by) the behavior
  - We need a domain expert for this step

- **Feature-Behavior Association:**
  - Find the relationship between features and behavior
  - E.g., use decision tree learning

- **Evaluation**:
  - The findings are due to the features and not to externalities.
  - E.g., we can use
    - classification accuracy
    - randomization tests (discussed later!)
    - or causality testing algorithms

Granger Causality

Granger Causality. Assume we are given two temporal variables $X = \{X_1, X_2, \ldots, X_t, X_{t+1}, \ldots\}$ and $Y = \{Y_1, Y_2, \ldots, Y_t, Y_{t+1}, \ldots\}$. Variable X "*Granger causes*" variable Y when historical values of X can help better predict Y than just using the historical values of Y.

Consider a linear regression model

- We can predict $\boldsymbol{Y_{t+1}}$ by using either $\boldsymbol{Y_{1,} Y_{2} \ldots Y_{t}}$ or a combination of $\boldsymbol{X_{1,} X_{2} \ldots X_{t}}$ and $\boldsymbol{Y_{1,} Y_{2} \ldots Y_{t}}$

$$Y_{t+1} = \sum_{i=1}^{t} a_i Y_i + \epsilon_1$$
$$Y_{t+1} = \sum_{i=1}^{t} a_i Y_i + \sum_{i=1}^{t} b_i X_i + \epsilon_2$$

- If $\boldsymbol{\varepsilon_2 < \varepsilon_1}$ then $\boldsymbol{X}$ Granger Causes $\boldsymbol{Y}$
  - **Why is this not causality?**

II. Individual Behavior Modeling

Individual Behavior Modeling

- Models in
  - Economics, Game Theory, and Network Science

We can use:

1. Threshold Models: we need to learn thresholds and weights

- W_{ij} can be defined as the fraction of times user i buys a product and user j buys the same product **soon** after that
  - When is soon?

– Similarly, thresholds can be estimated by taking into account the average number of friends who need to buy a product before user i decides to buy it.

– What if friends don't buy the same products?

- We can find the most similar individuals or items (similar to collaborative filtering methods)

2. Cascade Models

III. Individual Behavior Prediction

Individual Behavior Prediction

- Most behaviors result in newly formed links in social media.
  - It can be a link to a user, as in befriending behavior;
  - A link to an entity, as in buying behavior; or
  - A link to a community, as in joining behavior.

- We can formulate many of these behaviors as a **link prediction** problem.

Link Prediction - Setup

- Given a graph $G(V, E)$, let $e(u, v)$ denote edge between nodes u and v
  - $t(e)$ denotes the time that the edge was formed
- Let $G[t_1, t_2]$ represent the subgraph of G such that all edges are created between t_1 and t_2
  - i.e., for all edges e in this subgraph, $t_1 < t(e) < t_2$.
- Given four time stamps $t_{11} < t_{12} < t_{21} < t_{22}$ a link prediction algorithm is given
  - The subgraph $\mathrm{G}(t_{11}, t_{12})$ (**training interval**) and
  - Is expected to predict edges in $\mathrm{G}(t_{21}, t_{22})$ (**testing interval**).
- We can only predict edges for nodes that exist in the **training period**
- Let $\mathrm{G}(V_{train}, E_{train})$ be our **training graph**. Then, a link prediction algorithm generates a sorted list of most probable edges in

$$V_{train} \times V_{train} - E_{train}$$

Link Prediction - Algorithms

- Assign $\sigma(x, y)$ to every edge $e(x, y)$

- Edges sorted by this value in decreasing order will form our ranked list of predictions

- Any similarity measure between two nodes can be used for link prediction;
  - Network measures (**Chapter 3**) are useful here.

- We will review some well-known methods
  - Node Neighborhood-Based Methods
  - Path-Based Methods

Node Neighborhood-Based Methods

Node Neighborhood-Based Methods

- **Common Neighbors**:
  - The more common neighbors that two nodes share, the more similar they are

$$\sigma(x, y) = |N(x) \cap N(y)|$$

- **Jaccard Similarity:**
  - The likelihood of a node that is a neighbor of either x **or** y to be a common neighbor

$$\sigma(x, y) = \frac{|N(x) \cap N(y)|}{|N(x) \cup N(y)|}$$

Node Neighborhood-Based Methods

- **Adamic-Adar**:
  - If two individuals share a neighbor
  - and that neighbor is a **rare** neighbor,
  - it should have a higher impact on their similarity.

$$\sigma(x,y) = \sum_{z\in N(x)\cap N(y)} \frac{1}{\boxed{\log|N(z)|}}$$

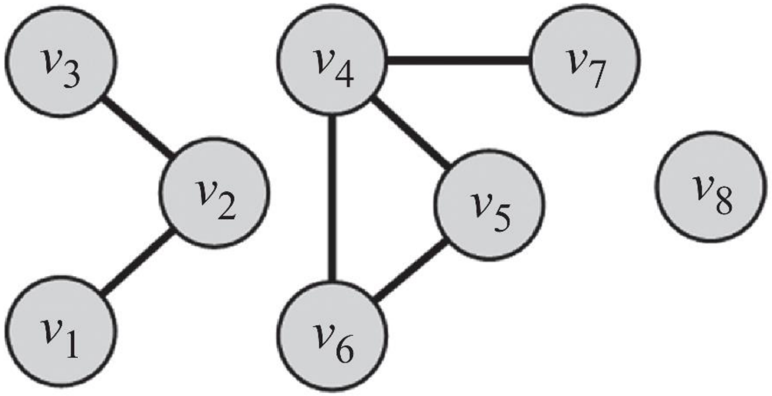
Rareness

- **Preferential Attachment**:
  - Nodes of higher degree have a higher chance of getting connected to incoming nodes

$$\sigma(x,y) = |N(x)|.|N(y)|$$

Example

Compute the edge score for edge (5,7)



$$
\begin{aligned}
(\textit{Common Neighbor})\ \sigma(5,7) &= |\{4,6\} \cap \{4\}| = 1 \\
(\textit{Jaccard})\ \sigma(5,7) &= \frac{|\{4,6\} \cap \{4\}|}{|\{4,6\} \cup \{4\}|} = \frac{1}{2} \\
(\textit{Adamic and Adar})\ \sigma(5,7) &= \frac{1}{\log|\{5,6,7\}|} = \frac{1}{\log 3} \\
(\textit{Preferential Attachment})\ \sigma(5,7) &= |\{4\}| \cdot |\{4,6\}| = 1 \times 2 = 2
\end{aligned}
$$

Path-Based Methods

Path-Based Measures

- **Katz Measure**:

$$\sigma(x,y) = \sum_{l=1}^{\infty} \beta^l |paths_{x,y}^{<l>}|$$

  - $|paths_{x,y}^{<l>}|$denotes the number of paths of length l between x and y
  - β is a constant that exponentially damps longer paths
    - When β is small, Katz measure reduces to common neighbor
  - It can be reformulated in closed form as

$$(I - \beta A)^{-1} - I$$

Path-Based Measures

- **Hitting Time and Commute Time**:
  - Consider a random walk that starts at node x and moves to adjacent nodes uniformly.
  - Hitting time is the expected number of random walk steps needed to reach y starting from x.
  - A smaller hitting time implies a higher similarity
    - A negation can turn it into a similarity measure

$$\sigma(x,y) = -H_{x,y}$$

  - **If y is highly connected random walks are more likely to visit y**
    - We can normalize it using the stationary probability

$$\sigma(x,y) = -H_{x,y}\pi_y$$

Path-Based Measures

  - **Hitting time** is not symmetric, we can use **commute time** instead, or its normalized version

$$\sigma(x, y) = -(H_{x,y} + H_{y,x})$$
$$\sigma(x, y) = -(H_{x,y}\pi_y + H_{y,x}\pi_x)$$

- **Rooted PageRank**: the stationary probability of y, when at each random walk run you can jump to x with probability P and to a random node with $1 - P$
- **SimRank**: Recursive definition of similarity

$$\sigma(x, y) = \gamma \cdot \frac{\sum_{x' \in N(x)} \sum_{y' \in N(y)} \sigma(x', y')}{|N(x)||N(y)|}$$

Link Prediction - Evaluation

- After one of the aforementioned measures is selected, a list of the top most similar pairs of nodes are selected.

- These pairs of nodes denote edges predicted to be the most likely to soon appear in the network.

- Performance (***precision***, ***recall***, or ***accuracy***) can be evaluated using the testing graph and by comparing the number of the testing graph's edges that the link prediction algorithm successfully reveals.

- Performance is usually **very low**, since many edges are created due to reasons not solely available in a social network graph.
  - **Solution:** a common baseline is to compare the performance with random edge predictors and report the **factor improvements** over random prediction.

Collective Behavior

Collective Behavior

- First Defined by sociologist Robert Park

- **Collective Behavior:** A group of individuals behaving in a similar way
- It can be planned and coordinated, but often is spontaneous and unplanned

Examples

- Individuals standing in line for a new product release
- Posting messages online to support a cause or to show support for an individual

I. Collective Behavior Analysis

Collective Behavior Analysis

- We can analyze collective behavior by analyzing individuals performing the behavior

- We can then put together the results of these analyses
  - The result would be the **expected behavior** for a large population

- **OR**, we can analyze the population as a whole
  - Not very popular for **analysis**, as individuals are ignored
  - Popular for **Prediction** purposes

Example – Analyzing User Migrations

Users migrate in social media due to their limited time and resources

- Sites are interested in keeping their users, because they are valuable assets that help contribute to their growth and generate revenue by increasing traffic

Two types of migrations:

- **Site migration**: For any user who is a member of two sites S_1 and S_2 at time t_i, and is only a member of S_2 at time $t_j > t_i$, then the user is said to have migrated from site S_1 to S_2.

- **Attention Migration:** For any user who is a member of two sites S_1 and S_2 and is active at both at time t_i, if the user becomes inactive on S_1 and remains active on S_2 at time $t_j > t_i$, then the user's attention is said to have migrated away from site S_1 and toward site S_2.

Collective Behavior Analysis - Example

- Activity (or inactivity) of a user can be determined by observing the user's actions performed on the site.

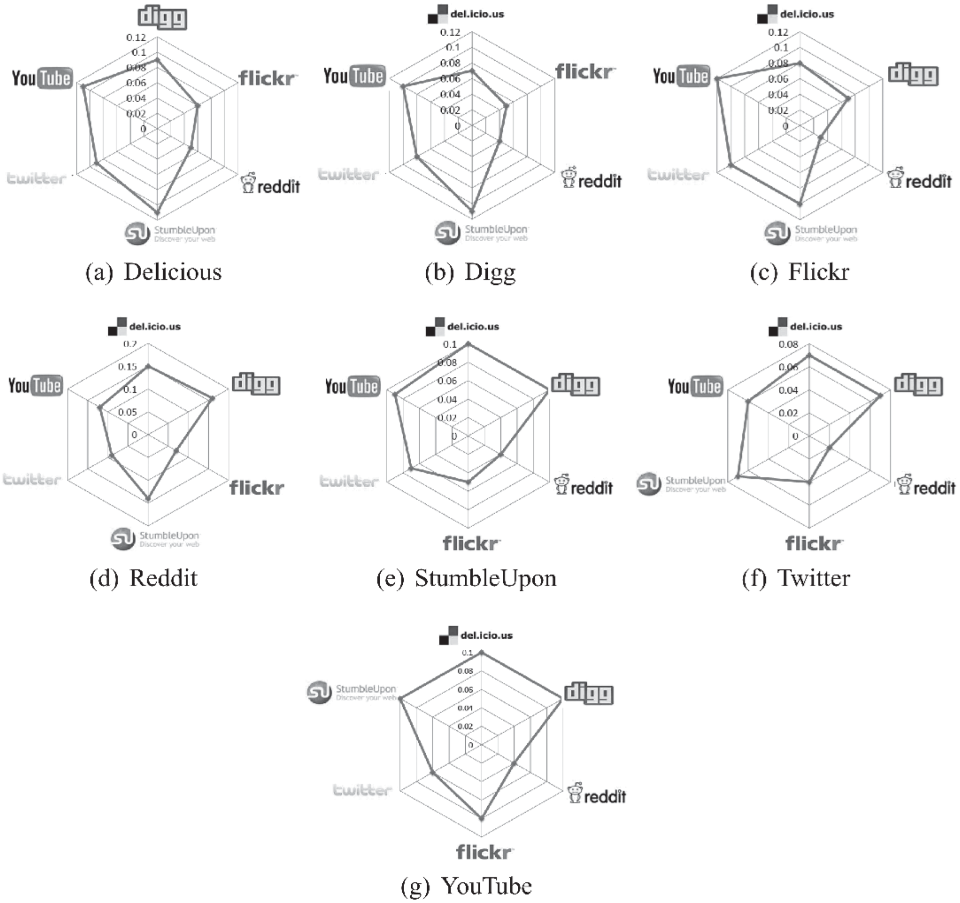
- We can consider a user active in $[t, t + X\ ]$, if the user has performed at least one action on the site during this period
  - Otherwise, the user is considered inactive.

- The interval could be measured at different granularity levels
  - E.g., days, weeks, months, and years.
  - It is common to set $X\ =\ 1$ month.

- We can analyze migrations of **individuals** and then measure the rate at which the **populations** are migrating across sites.
  - We can use the methodology for individual behavior analysis

The Observable Behavior

- Sites migration is rarely observed
- Attention migration is clearly observable
- We need to take multiple steps to observe it:
  - Users are required to be identified on multiple networks (challenging!)
    - Some ideas: **John.Smith1** on Facebook is **JohnSmith** on Twitter



(a) Delicious

(b) Digg

(c) Flickr

(d) Reddit

(e) StumbleUpon

(f) Twitter

(g) YouTube

Features

- **User Activity**: more active users are less likely to migrate
  - e.g., number of tweets, posts, or photos
- **User Network Size:** a user with more social ties (i.e., friends) in a social network is less likely to move
  - e.g., number of friends
- **User Rank:** a user with high status in a network is less likely to move to a new one where he or she must spend more time getting established.
  - e.g., centrality scores
  - External rank: your citations, how many have referred to your article, …

Feature-Behavior Association

- Given two snapshots of a network, we know if users migrated or not.
- Let vector $\mathrm{Y} \in \mathbb{R}^{n}$ indicate whether any of our n users have migrated or not.
- Let $X_t \in \mathbb{R}^{3\times n}$ be the features collected (activity, friends, rank) for any one of these users at time stamp t.
- The correlation between features X and labels Y can be computed via logistic regression.
- How can we verify that this correlation is not random?

Evaluation Strategy

To verify if the correlation between features and the migration behavior is not random

- We can construct a random set of migrating users
  - compute X_{Random} and $Y_{Random}$for them
- Find the correlation between these random variables (e.g., regression coefficients) and it should be significantly different from what we obtained using real-world observations

We can use χ^2 (Chi-square) test for significance testing

From Original Dataset

From Random Dataset

$$\chi^2 = \sum_{i=1}^{n} \frac{(A_i - R_i)^2}{R_i}$$

II. Collective Behavior Modeling

Collective Behavior Modeling

- Collective behavior can be conveniently modeled using some of the techniques discussed in **Chapter 4 - Network Models.**
- We want models that can mimic characteristics observable in the population.
- In network models, node properties rarely play a role
  - Reasonable for modeling collective behavior.

III. Collective Behavior Prediction

Collective Behavior Prediction

- From previous chapters, we could use
  - Linear Influence Model (LIM)
  - Epidemic Models

- Collective behavior can be analyzed either in terms of
  1. individuals performing the collective behavior or
  2. based on the population as a whole. (**More Common**)

- When predicting collective behavior,
  - We are interested in predicting the intensity of a phenomenon, which is due to the collective behavior of the population
  - e.g., how many of them will vote?

- We can utilize a data mining approach where features that describe the population well are used to predict a response variable
  - i.e., the intensity of the phenomenon

- A **training-testing** framework or correlation analysis is used to determine the generalization and the accuracy of the predictions.

Predicting Box Office Revenue for Movies

1. Set the target variable that is being predicted
   - **In our example:** the revenue that a movie produces.
   - The revenue is the direct result of the collective behavior of going to the theater to watch the movie.

2. Identify features in the population that may affect the target variable
   - the average hourly number of tweets related to the movie for each of the seven days prior to the movie opening (seven features)
   - The number of opening theaters for the movie (one feature).

3. Predict the target variable using a supervised learning approach, utilizing the features determined in step 2.

4. Measure performance using supervised learning evaluation.

The predictions using this approach are closer to reality than that of the Hollywood Stock Exchange (HSX), which is the gold standard for predicting revenues for movies

Generalizing the Idea

- **Target variable y**
  - Some feature A that quantifies the attention
  - Some feature P that quantifies the publicity

- Train a regression model

$$y = w_1 A + w_2 P + \epsilon$$