Data Mining Based Social Network Analysis
**Nodes**: individuals

**Links**: social relationship (family/work/friendship/etc.)

*S. Milgram (1967)*

**Six Degrees of Separation**

Social networks: Many individuals with diverse social interactions between them.

*John Guare*
Small world phenomenon: Milgram’s experiment

Instructions:
Given a target individual (stockbroker in Boston), pass the message to a person you correspond with who is “closest” to the target.
Small world phenomenon: 
Milgram’s experiment

Outcome:
20% of initiated chains reached target
average chain length = 6.5

“Six degrees of separation”
Small world phenomenon: Milgram’s experiment repeated

email experiment

• 18 targets
• 13 different countries

• 60,000+ participants
• 24,163 message chains
• 384 reached their targets
• average path length 4.0

World Wide Web

**Nodes**: WWW documents

**Links**: URL links

800 million documents
(S. Lawrence, 1999)


**ROBOT**: collects all URL’s found in a document and follows them recursively
World Wide Web

Expected Result

\[ \langle k \rangle \sim 6 \]
\[ P(k=500) \sim 10^{-99} \]
\[ N_{WWW} \sim 10^9 \]
\[ \Rightarrow N(k=500) \sim 10^{-90} \]

Real Result

\[ P_{out}(k) \sim k^{-\gamma_{out}} \]
\[ \gamma_{out} = 2.45 \]

\[ P_{in}(k) \sim k^{-\gamma_{in}} \]
\[ \gamma_{in} = 2.1 \]

What does that mean?

Poisson distribution

Power-law distribution

Exponential Network

Scale-free Network
Scale-Free Networks

- Preferential attachment explains
  - heavy-tailed degree distributions
  - small diameter (~log(N), via “hubs”)

- Will *not* generate high clustering coefficient
  - no bias towards local connectivity, but towards hubs
Detecting Fraud

Which transactions are likely to be fraudulent?
Modeling Epidemics

http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0040961
Which link is missing?

(A) vanilla extract, celery
(B) pepper, onion

Recipe recommendation using ingredient networks. Teng et al., 2012.
A network is a collection of objects where some pairs of objects are connected by links.

What is the structure of the network?
- **Objects**: nodes, vertices
- **Interactions**: links, edges
- **System**: network, graph

\[ G(N,E) \]
Network vs Graph

- **Network** often refers to real systems
  - Web, Social network, Metabolic network
  
  **Language:** Network, node, link

- **Graph** is a mathematical representation of a network
  - Web graph, Social graph (a Facebook term)
  
  **Language:** Graph, vertex, edge
- If you connect individuals that work with each other, you will explore a **professional network**

- If you connect scientific papers that cite each other, you will be studying the **citation network**

- If you connect all papers with the same word in the title, you will be exploring what? It is a network, nevertheless
How to define a network?

- What are nodes?
- What are edges?

**Choice of the proper network representation of a given domain/problem determines our ability to use networks successfully:**

- In some cases there is a unique, unambiguous representation
- In other cases, the representation is by no means unique
- The way you assign links will determine the nature of the question you can study
Introduction to Social Network Analysis
Social Networks

- A **social network** is a social structure of people, related (directly or indirectly) to each other through a common relation or interest.

- **Social network analysis (SNA)** is the study of social networks to understand their structure and behavior.

(Source: Freeman, 2000)
Framework for Social Network Analysis
Types of Social Network Analysis

- **Sociocentric (whole) network analysis**
  - Emerged in sociology
  - Involves quantification of interaction among a socially well-defined group of people
  - Focus on identifying global structural patterns
  - Most SNA research in organizations concentrates on sociometric approach

- **Egocentric (personal) network analysis**
  - Emerged in anthropology and psychology
  - Involves quantification of interactions between an individual (called *ego*) and all other persons (called *alters*) related (directly or indirectly) to ego
  - Make generalizations of features found in personal networks
  - Difficult to collect data, so till now studies have been rare
Social science networks have widespread application in various fields.

Most of the analyses techniques have come from Sociology, Statistics and Mathematics.

See (Wasserman and Faust, 1994) for a comprehensive introduction to social network analysis.

Classification based on Contractor 2006.
Networks Online

**Communication networks:**
- Intrusion detection, fraud
- Churn prediction

**Social networks:**
- Link prediction, friend recommendation
- Social circle detection, community detection
- Social recommendations
- Identifying influential nodes, Information virality

**Information networks:**
- Navigational aids
Data Mining for Social Network Analysis
DM for SNA

- Cascading Behaviour, Influence propagation, Influence Diffusion
- Community Detection and Extraction
- Link Mining and Prediction

- Search in Social Networks
- Trust in Social Networks
- Anonymity in Social Networks
- Other Research
Cascading Behavior
Influence Propagation
Influence Diffusion
Information cascade in a social network
Can cascades be predicted? Cheng et al., WWW ’14.
60-90% of LinkedIn users signed up due to an invitation from another user. Global Diffusion via Cascading Invitations: Structure, Growth, and Homophily. Anderson et al., WWW ‘15.
Cascading Models

- **Model of Diffusion of Innovation (Young, 2000)**
  - Interactions between the agents are weighted
  - Directed edges represent influence of one agent on the other
  - Agents have to choose between outcomes
    - The choice is based on a utility function which has an individual and a social component
  - The social component depends upon the choices made by the neighbours
  - Under the assumption of a logistic response, diffusion time is independent of number of actors and initial state and a final stable state will be reached
The Spread of Obesity in a Large Social Network over 32 Years

Data set: 12,067 people from 1971 to 2003, 50K links

<table>
<thead>
<tr>
<th>Alter Type</th>
<th>Increase in Risk of Obesity in Ego (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ego-perceived friend</td>
<td>0</td>
</tr>
<tr>
<td>Mutual friend</td>
<td>0</td>
</tr>
<tr>
<td>Alter-perceived friend</td>
<td>0</td>
</tr>
<tr>
<td>Same-sex friend</td>
<td>0</td>
</tr>
<tr>
<td>Opposite-sex friend</td>
<td>0</td>
</tr>
<tr>
<td>Spouse</td>
<td>0</td>
</tr>
<tr>
<td>Sibling</td>
<td>0</td>
</tr>
<tr>
<td>Same-sex sibling</td>
<td>0</td>
</tr>
<tr>
<td>Opposite-sex sibling</td>
<td>0</td>
</tr>
<tr>
<td>Immediate neighbor</td>
<td>0</td>
</tr>
</tbody>
</table>

Obese Friend → 57% increase in chances of obesity
Obese Sibling → 40% increase in chances of obesity
Obese Spouse → 37% increase in chances of obesity
Network contagion: other examples

How your friends’ friends’ friends’ affect everything you feel, think, and do

Christakis and Fowler

back pain (spread from West Germany to East Germany after the fall of the Berlin Wall)
suicide (well known to spread throughout communities on occasion)
politics (the denser your network of connections, the more ideologically intense your beliefs)
Influence or Homophily?

Homophily

tendency to stay together with people similar to you

“Birds of a feather flock together”

Social influence

a force that person A (i.e., the influencer) exerts on person B to introduce a change of the behavior and/or opinion of B

Influence is a causal process

RP#1: How to distinguish social influence from homophily and other external factors
See e.g.,
Crandall et al. (KDD’08) “Feedback Effects between Similarity and Social Influence in Online Communities”
Anagnostopoulos et al. (KDD’08) “Influence and correlation in social networks”
Influence in on-line social networks

Users perform actions:
- post messages, pictures, video
- buy, comment, link, rate, share, like, retweet

Users are connected with other users:
- interact, influence each other

Actions propagate
IDEA: exploit social influence for marketing

Basic assumption: word-of-mouth effect, thanks to which actions, opinions, buying behaviors, innovations and so on, propagate in a social network.

Target users who are likely to invoke word-of-mouth diffusion, thus leading to additional reach, clicks, conversions, or brand awareness

Target the influencers

Sharing and social influence

How frequently do you share recommendations online?

- Every few weeks: 19%
- Every few months: 42%
- Every few days: 8%
- Daily: 2%
- Never: 29%
Bring me the influencers!

Influencers increase brand awareness/product conversion through WOMM
Influencers advocate brand
Influencers influence a purchasing action from their peers

Some of the many startups involved in social influence

Klout (http://klout.com)
Measure of overall influence online (mostly twitter, now FB and linkedin)
Score = function of true reach, amplification probability and network influence
Claims score to be highly correlated to clicks, comments and retweets

Peer Index (http://www.peerindex.net)
Identifies/Scores authorities on the social web by topic

SocialMatica (http://www.socialmatica.com)
Ranks 32M people by vertical/topic, claims to take into account quality of authored content

Influencer50 (http://www.influencer50.com)
Clients: IBM, Microsoft, SAP, Oracle and a long list of tech companies
Svnetwork, Bluecalypso, CrowdBooster, Sproutsocial, TwentyFeet, EmpireAvenue, Twitaholic
(there’s more ... )
Viral Marketing and Influence Maximization

**Business goal (Viral Marketing):** exploit the “word-of-mouth” effect in a social network to achieve marketing objectives through self-replicating viral processes

**Mining problem statement (Influence Maximization):** find a seed-set of influential people such that by targeting them we maximize the spread of viral propagations

Hot topic in Data Mining research since 10 years:
Domingos and Richardson *Mining the network value of customers* (KDD’01)
Domingos and Richardson *Mining knowledge-sharing sites for viral marketing* (KDD’02)
Kempe et al. *Maximizing the spread of influence through a social network* (KDD’03)
Information Diffusion

• In February 2013, during the third quarter of Super Bowl XLVII, a power outage stopped the game for 34 minutes. Oreo, a sandwich cookie company, tweeted during the outage: “Power out? No Problem, You can still dunk it in the dark”. The tweet caught on almost immediately, reaching nearly 15,000 retweets and 20,000 likes on Facebook in less than 2 days.

• A simple tweet diffused into a large population of individuals. It helped the company gain fame with minimum budget in an environment where companies spent as much as 4 million dollars to run a 30 second ad during the super bowl.

• This is an example of Information Diffusion.

Information diffusion: process by which a piece of information (knowledge) is spread and reaches individuals through interactions.
Information Diffusion

• **Sender(s).** A sender or a small set of senders that initiate the information diffusion process;

• **Receiver(s).** A receiver or a set of receivers that receive diffused information. Commonly, the set of receivers is much larger than the set of senders and can overlap with the set of senders; and

• **Medium.** This is the medium through which the diffusion takes place. For example, when a rumor is spreading, the medium can be the personal communication between individuals
Information Diffusion Types

We define the process of interfering with information diffusion by expediting, delaying, or even stopping diffusion as Intervention.
Herd Behavior

- Network is observable
- Only public information is available
Herd Behavior Example

• Consider people participating in an online auction.

• In this case, individuals can observe the behavior of others by monitoring the bids that are being placed on different items.
• Individuals are connected via the auction’s site where they can not only observe the bidding behaviors of others, but can also often view profiles of others to get a feel for their reputation and expertise.

• In these online auctions, it is common to observe individuals participating actively in auctions, where the item being sold might otherwise be considered unpopular.

• This is due to individuals trusting others and assuming that the high number of bids that the item has received is a strong signal of its value. In this case, Herd Behavior has taken place.
Herd Behavior: Popular Restaurant Experiment

• Assume you are on a trip in a metropolitan area that you are less familiar with.

• Planning for dinner, you find restaurant A with excellent reviews online and decide to go there.

• When arriving at A, you see A is almost empty and restaurant B, which is next door and serves the same cuisine, almost full.

• Deciding to go to B, based on the belief that other diners have also had the chance of going to A, is an example of herd behavior.
Herding: Elevator Example

http://www.youtube.com/watch?v=zNNz0yzHcwg
Herd Behavior

Herd behavior describes when a group of individuals performs actions that are highly correlated without any plans.

Main Components of Herd Behavior

• A method to transfer behavior among individuals or to observe their behavior
• A connection between individuals
Network Observability in Herb Behavior

In herd behavior, individuals make decisions by observing all other individuals’ decisions

• In general, herd behavior’s network is close to a complete graph where nodes can observe at least most other nodes and they can observe public information
  • For example, they can see the crowd
Herding: Urn Experiment

• There is an urn in a large class with three marbles in it

![Marble Configuration](image)

*50% 50%*

• During the experiment, each student comes to the urn, picks one marble, and checks its color in private.
• The student predicts majority blue or red, writes her prediction on the blackboard, and puts the marble back in the urn.
• Students can’t see the color of the marble taken out and can only see the predictions made by different students regarding the majority color on the board
Urn Experiment: First and Second Student

• First Student:
  • *Board*: -
    • Observed: B $\rightarrow$ Guess: B
      - or -
    • Observed: R $\rightarrow$ Guess: R

• Second Student:
  • *Board*: B
    • Observed: B $\rightarrow$ Guess: B
      - or -
    • Observed: R $\rightarrow$ Guess: R/B (flip a coin)
Urn Experiment: Third Student

• *If board: B, R*
  • Observed: B → Guess: B, or
  • Observed: R → Guess: R

• *If board: B, B*
  • Observed: B → Guess: B, or
  • Observed: R → Guess: B (Herding Behavior)

The forth student and onward

• Board: B,B,B
  • Observed: B/R → Guess: B
Urn Experiment
Herding Intervention

In herding, the society only has access to public information.

Herding may be intervened by releasing private information which was not accessible before.

The little boy in “The Emperor’s New Clothes” story intervenes the herd by shouting “he's got no clothes on”
Leaders and Tribes: a pattern mining approach*

Given a time threshold $\tau$, in a given propagation, define the followers of a user $u$, those ones in the “subtree” of $u$, that activate within $\tau$ from $u$.

A user is a leader w.r.t. a given action when the number of his followers exceeds a given threshold.

Tribe Leaders:
Previous definition does not force the set of followers for different actions to be the same. If we add this constraint we obtain tribe leaders.

A user to be identified as a leader must act as such sufficiently often, i.e., for a number of actions larger than a given threshold.

Additional constraints:
- Confidence
- Genuineness

Develop efficient algorithms that make only a pass over the actions log.

*Goyal, Bonchi, Lakshmanan (CIKM’08) “Discovering Leaders from Community Actions” (ICDE’09, demo) “GuruMine: a Pattern Mining System for Discovering Leaders and Tribes”
Discovering Leaders from Community Actions

Figure 1: (a) Example social graph; (b) A log of actions; (c) Propagation of action a and (d) of action b.
Information Cascade

- In the presence of a network
- Only local information is available
Information Cascade

• In social media, individuals commonly repost content posted by others in the network. This content is often received via immediate neighbors (friends).

• An **Information Cascade occurs as information propagates** through friends

• An information cascade is defined as a piece of information or decision being cascaded among a set of individuals, where
  • 1) individuals are connected by a network and
  • 2) individuals are only observing decisions of their immediate neighbors (friends).

• Therefore, cascade users have less information available to them compared to herding users, where almost all information about decisions are available.

In cascading, local information is available to the users, but in herding the information about the population is available.
Underlying Assumptions for Cascade Models

• The network is represented using a directed graph. Nodes are actors and edges depict the communication channels between them. A node can only influence nodes that it is connected to;

• Decisions are binary - nodes can be either active or inactive. An active nodes means that the node decided to adopt the behavior, innovation, or decision;

• A node, once activated, can activate its neighboring nodes; and

• Activation is a progressive process, where nodes change from inactive to active, but not vice versa 1.
Independent Cascade Model (ICM)

• **Independent Cascade Model** is a sender centric model of cascade
  • In this model each node has one chance to activate its neighbors

• Considering nodes that are active as senders and nodes that are being activated as receivers,
  • The *linear threshold model* concentrates on the receiver (to be discussed later).
  • The independent cascade model model concentrates on the sender
Independent Cascade Model (IC)

Every arc \((u,v)\) has associated the probability \(p(u,v)\) of \(u\) influencing \(v\)

Time proceeds in discrete steps

At time \(t\), nodes that became active at \(t-1\) try to activate their inactive neighbors, and succeed according to \(p(u,v)\)
Independent Cascade Model (IC)

Every arc \((u,v)\) has associated the probability \(p(u,v)\) of \(u\) influencing \(v\).

Time proceeds in discrete steps.

At time \(t\), nodes that became active at \(t-1\) try to activate their inactive neighbors, and succeed according to \(p(u,v)\).
Independent Cascade Model (IC)

Every arc \((u,v)\) has associated the probability \(p(u,v)\) of \(u\) influencing \(v\).

Time proceeds in discrete steps.

At time \(t\), nodes that became active at \(t-1\) try to activate their inactive neighbors, and succeed according to \(p(u,v)\).
Independent Cascade Model (IC)

Every arc \((u,v)\) has associated the probability \(p(u,v)\) of \(u\) influencing \(v\). Time proceeds in discrete steps. At time \(t\), nodes that became active at \(t-1\) try to activate their inactive neighbors, and succeed according to \(p(u,v)\)
Algorithm 7.1 Independent Cascade Model (ICM)

Require: Diffusion graph $G(V, E)$, set of initial activated nodes $A_0$, activation probabilities $p_{v,w}$
1: return Final set of activated nodes $A_\infty$
2: $i = 0$;
3: while $A_i \neq \emptyset$ do
4: 
5: $i = i + 1$;
6: $A_i = \emptyset$;
7: for all $v \in A_{i-1}$ do
8: for all $w$ neighbor of $v$, $w \notin \cup_{j=0}^{i-1} A_j$ do
9: \hspace{1cm} rand = generate a random number in [0,1];
10: \hspace{1cm} if rand $< p_{v,w}$ then
11: \hspace{1cm} \hspace{1cm} activate $w$;
12: \hspace{1cm} \hspace{1cm} $A_i = A_i \cup \{w\}$;
13: \hspace{1cm} end if
14: \hspace{1cm} end for
15: end for
16: end while
17: $A_\infty = \cup_{i=0}^{\infty} A_j$;
18: Return $A_\infty$;
Independent Cascade Model Diffusion Process

Step 0

Step 1

Step 2

Step 3

Step 4

Final Stage
How should we organize revolt?

• You live in an oppressive society
• You know of a demonstration against the government planned tomorrow
• If a lot of people show up, the government will fall
• If only a few people show up, the demonstrators will be arrested and it would have been better had everyone stayed at home
Pluralistic Ignorance

• You should do something if you believe you are in the majority!
• Dictator tip: Pluralistic ignorance – erroneous estimates about the prevalence of certain opinions in the population.
Organizing the Revolt: The Model

- Personal threshold $k$: “I will show up if am sure at least $k$ people in total (including myself) will show up”
- Each node only knows the thresholds and attitudes of all their direct friends.
- Can we predict if a revolt can happened based on the network structure?
Which Network Will Revolt?
Linear Threshold Model

An actor would take an action if the number of his friends who have taken the action exceeds (reaches) a certain threshold

- Each node $v$ chooses a threshold $\Theta_v$ randomly from a uniform distribution in an interval between 0 and 1.
- In each discrete step, all nodes that were active in the previous step remain active.
- The nodes satisfying the following condition will be activated

$$\sum_{w \in N_v, w \text{ is active}} b_{w,v} \geq \Theta_v$$
Linear Threshold Model (LT)

- Every arc (u,v) has associated a weight \( b(u,v) \) such that the sum of incoming weights in each node is \( \leq 1 \)
- Time proceeds in discrete steps
- Each node \( v \) picks a random threshold \( \theta_v \sim U[0,1] \)
- A node \( v \) becomes active when the sum of incoming weights from active neighbors reaches \( \theta_v \)
Linear Threshold Model (LT)

- Every arc \((u,v)\) has associated a weight \(b(u,v)\) such that the sum of incoming weights in each node is \(\leq 1\).
- **Time** proceeds in discrete steps.
- Each node \(v\) picks a random threshold \(\theta_v \sim U[0,1]\).
- A node \(v\) becomes active when the sum of incoming weights from active neighbors reaches \(\theta_v\).
Linear Threshold Model (LT)

- Every arc \((u,v)\) has associated a weight \(b(u,v)\) such that the sum of incoming weights in each node is \(\leq 1\)
- Time proceeds in discrete steps
- Each node \(v\) picks a random threshold \(\theta_v \sim U[0,1]\)
- A node \(v\) becomes active when the sum of incoming weights from active neighbors reaches \(\theta_v\)
Linear Threshold Diffusion Process

Step 0

Step 1

Step 2

Step 3

Final Stage
Maximizing the spread of cascades

• **Maximizing the Spread of Cascades** is the problem of finding a small set of nodes in a social network such that their aggregated spread in the network is maximized

• **Applications**
  • Product marketing
  • Influence
Problem Setting

• Given
  • A limited budget $B$ for initial advertising (e.g., give away free samples of product)
  • Estimating spread between individuals

• Goal
  • To trigger a large spread (e.g., further adoptions of a product)

• Question
  • Which set of individuals should be targeted at the very beginning?
Maximizing the Spread of Cascade: Example

• We need to pick k nodes such that maximum number of nodes are activated
Maximizing the Spread of Cascade

Select one seed

Select two seeds
Problem Statement

• Spread of node set S: f(S)
  • An expected number of active nodes, if set S is the initial active set

• Problem:
  • Given a parameter \( k \) (budget), find a \( k \)-node set S to maximize \( f(S) \)
  • A constrained optimization problem with \( f(S) \) as the objective function
Some Facts Regarding this Problem

• Bad News
  • For a submodular function monotone non-negative \( f \), finding a \( k \)-element set \( S \) for which \( f(S) \) is maximized is an NP-hard optimization problem
  • It is NP-hard to determine the optimum for influence maximization for both independent cascade model and linear threshold model.

• Good News
  • We can use Greedy Algorithm
    • Start with an empty set \( S \)
    • For \( k \) iterations:
      Add node \( v \) to \( S \) that maximizes \( f(S + v) - f(S) \).
  • How good (or bad) it is?
    • Theorem: The greedy algorithm is a \((1 − 1/e)\) approximation.
    • The resulting set \( S \) activates at least \((1- 1/e) > 63\%\) of the number of nodes that any size-\( k \) set \( S \) could activate.
Cascade Maximization: A Greedy approach

Maximizing the cascade is a NP-hard problem but it is proved that the greedy approaches gives a solution that is at least 63 % of the optimal.

Given a network and a parameter $k$, which $k$ nodes should be selected to be in the activation set $B$ in order to maximize the cascade in terms of the total number of active nodes?

• Let $\sigma(B)$ denote the expected number of nodes that can be activated by $B$, the optimization problem can be formulated as follows:

$$\max_{B \subseteq V} \sigma(B) \text{ s.t. } |B| \leq k$$
Cascade Maximization: A Greedy Approach

Bad news: **NP-hard** optimization problem for both IC and LT models

Good news: we can use **Greedy algorithm**

**Algorithm 7.2** Maximizing the spread of cascades – Greedy algorithm

```
Require: Diffusion graph $G(V,E)$, budget $k$
1: return Seed set $S$ (set of initially activated nodes)
2: $i = 0$
3: $S = {}$
4: while $i < k$ do
5: $v = \arg \max_{v \in V \setminus S} f(S \cup \{v\})$
6: $S = S \cup \{v\}$
7: $i = i + 1$
8: end while
9: Return $S$
```

$\sigma_M(S)$ is **monotone** and **submodular**

**Theorem**: The resulting set $S$ activates at least $(1 - 1/e) > 63\%$ of the number of nodes that any size-$k$ set could activate

*Nemhauser et al. “An analysis of approximations for maximizing submodular set functions – (i)” (1978)*
Diffusion of Innovations

- The network is not observable
- Only public information is observable
Diffusion Curves

• Basis for models:
  • Probability of adopting new behavior depends on the number of friends who already adopted
• What is the dependence?

• Different shapes has consequences for models of diffusion
Real World Diffusion Curves

- DVD recommendation and LiveJournal community membership
Diffusion of Innovation

• an innovation is “an idea, practice, or object that is perceived as new by an individual or other unit of adoption”

• The theory of diffusion of innovations aims to answer why and how these innovations spread. It also describes the reasons behind the diffusion process, individuals involved, as well as the rate at which ideas spread.
Diffusion of Innovations Models

• First model was introduced by Gabriel Tarde in the early 20th century
Two-Step (multiple-step) Flow Model of Diffusion

• According to the two-step flow model, most information comes from mass media, which is then directed toward influential figures called opinion leaders.

• These leaders then convey the information (or form opinions) and act as hubs for other members of the society.
Modeling Diffusion of Innovations

This diffusion of innovation model describes the rate at which the number of adopters changes in terms of time:

\[
\frac{dA(t)}{dt} = i(t)[P - A(t)]
\]

- \(A(t)\) is the total population that adopted the innovation
- \(i(t)\) denotes the coefficient of diffusion corresponding to the innovativeness of the product being adopted
- \(P\) is the total number of potential adopters (till time \(t\))

- The rate depends on how innovative the product is
- The rate affects the potential adopters that have not yet adopted the product.
Information Diffusion: Mathematical Model

\[
\frac{dA(t)}{dt} = i(t)[P - A(t)] \quad \Rightarrow \quad A(t) = \int_{t_0}^{t} a(t)dt,
\]

the adopters at time \( t \)

Defining the diffusion coefficient by defining \( i(t) \) as a function of number of adopters \( A(t) \), \( (A_0: \text{the number of adopters at time } t_0) \)

\[
i(t) = \alpha + \alpha_0A_0 + \ldots + \alpha_tA(t) = \alpha + \sum_{i=t_0}^{t} \alpha_iA(i)
\]
Diffusion Models

Three models of diffusion:

\[
\frac{dA(t)}{dt} = i(t)[P - A(t)]
\]

- \(i(t) = \alpha\), \(\text{External-Influence Model}\)
- \(i(t) = \beta A(t)\), \(\text{Internal-Influence Model}\)
- \(i(t) = \alpha + \beta A(t)\), \(\text{Mixed-Influence Model}\)

- \(\alpha\): Innovativeness factor of the product
- \(\beta\): Imitation factor
Epidemics
Epidemics describes the process by which diseases spread. This process consists of:

- A pathogen (the disease being spread),
- A population of hosts (humans, animals, plants, etc.),
- A spreading mechanism (breathing, drinking, sexual activity, etc.)
Comparing Epidemics and Cascades

• Unlike information cascades and herding and similar to diffusion of innovations models, epidemic models assume an implicit network and unknown connections between individuals.

• This makes epidemic models more suitable when we are interested in global patterns, such as trends and ratios of people getting infected, and not in who infects whom.
How to Analyze Epidemics?

• Contact Network
  • look at how hosts contact each other and devise methods that describe how epidemics happen in networks.
  • A contact network is a graph where nodes represent the hosts and edges represent the interactions between these hosts. For instance, in the case of the HIV/AIDS, edges represent sexual interactions, and in the case of influenza, nodes that are connected represent hosts that breathe the same air.

• Fully-mixed
  • Analyze only the rates at which hosts get infected, recover, etc. and avoid considering network information

The models discussed here will assume:
• No contact network information is available
• The process by which hosts get infected is unknown
Models of Infection (Virus Propagation)

• How do virus/rumors propagate?
• Will a flu-like virus linger or will it die out soon?
• (Virus) birth rate $\beta$: probability that an infected neighbor attacks
• (Virus) death rate $\delta$: probability that an infected neighbor recovers
SI Model: Definition

• **Susceptible**
  • When an individual is in the susceptible state, he or she can potentially get infected by the disease.

• **Infected**
  • An infected individual has the chance of infecting susceptible parties
General Schemes
Susceptible-Infected-Recovered (SIR) Model

• Process:
  • Initially, some nodes are in the $I$ state and all others in the $S$ state.
  • Each node $v$ in the $I$ state remains infectious for a fixed number of steps $t$.
  • During each of the $t$ steps, node $v$ can infect each of its susceptible neighbors with probability $p$.
  • After $t$ steps, $v$ is no longer infectious or susceptible to further infections and enters state $R$.

• SIR is suitable for modeling a disease that each individual can only catches once during their life time.
Example SIR epidemic (t=1)
Susceptible-Infected-Susceptible (SIS) Model

- Cured nodes immediately become susceptible again.
- Virus “strength”: $s = \frac{\beta}{\delta}$
Example SIS Epidemic

Figure 21.5. In an SIS epidemic, nodes can be infected, recover, and then be infected again. In each step, the nodes in the infectious state are shaded.
Connection between SIS and SIR

• SIS model with t=1 can be represented as an SIR model by creating a separate copy of each node for each time step.
Question: Epidemic Threshold

• The epidemic (let's say a zombie plague) threshold of a graph is a value of $\tau$, such that
  • If strength $s = \frac{\beta}{\delta} < \tau$, then an epidemic can not happen.
• What should $\tau$ depend on?
  • Avg. degree? And/or highest degree?
  • And/or variance of degree?
  • And/or diameter?
Epidemic threshold in SIS model

- We have no epidemic if:

\[ \frac{\beta}{\delta} < \tau = \frac{1}{\lambda_{1,A}} \]

- Death rate
- Birth rate
- Epidemic threshold
- Largest eigenvalue of adjacency matrix A
Simulation Studies:

$\beta/\delta > \tau$

(above threshold)

$\beta/\delta = \tau$

(at the threshold)

$\beta/\delta < \tau$

(below threshold)

10,900 nodes and 31,180 edges
Experiments:

Does it matter how many people are initially infected?

(a) Below the threshold, $s=0.912$

(b) At the threshold, $s=1.003$

(c) Above the threshold, $s=1.1$
Epidemic Intervention

• Suppose that we have a susceptible society and want to prevent more spread by vaccinating the most vulnerable individuals

• How to find the most vulnerable individuals?

Randomly pick some nodes and ask them who is the most vulnerable from their point of view, then vaccinate those individuals!
Community Detection or Extraction
Communities

- **Community**: “subsets of actors among whom there are relatively strong, direct, intense, frequent or positive ties.”
  
  -- Wasserman and Faust, *Social Network Analysis, Methods and Applications*

- Community is a set of actors interacting with each other *frequently*

- A set of people without interaction is NOT a community
  - e.g. people waiting for a bus at station but don’t talk to each other
Community Detection

- **Community Detection**: “formalize the strong social groups based on the social network properties”

- Some social media sites allow people to join groups
  - Not all sites provide community platform
  - Not all people join groups

- Network interaction provides rich information about the relationship between users
  - Is it necessary to extract groups based on network topology?
  - Groups are *implicitly* formed
  - Can complement other kinds of information
  - Provide basic information for other tasks
Community Detection

- Community detection
  - a.k.a. grouping, clustering, finding cohesive subgroups
  - Given: a social network
  - Output: community membership of (some) actors

- Applications
  - Understanding the interactions between people
  - Visualizing and navigating huge networks
  - Forming the basis for other tasks such as data mining
Visualization after Grouping

4 Groups:
{}{1,2,3,5}
{}{4,8,10,12}
{}{6,7,11}
{}{9,13}

(Nodes colored by Community Membership)
Classification

• User Preference or Behavior can be represented as class labels
  • Whether or not clicking on an ad
  • Whether or not interested in certain topics
  • Subscribed to certain political views
  • Like/Dislike a product

• Given
  • A social network
  • Labels of some actors in the network

• Output
  • Labels of remaining actors in the network
Visualization after Prediction

**: Smoking**
**: Non-Smoking**
**: ? Unknown**

Predictions
6: Non-Smoking
7: Non-Smoking
8: Smoking
9: Non-Smoking
10: Smoking
11: Non-Smoking
12: Non-Smoking
Principles of Community Detection
Subjectivity of Community Definition

A densely-knit community

Each component is a community

Definition of a community can be subjective.
Taxonomy of Community Criteria

• Criteria vary depending on the tasks

• Roughly, community detection methods can be divided into 4 categories (not exclusive):

  • **Node-Centric Community**
    • Each node in a group satisfies certain properties

  • **Group-Centric Community**
    • Consider the connections within a group as a whole. The group has to satisfy certain properties without zooming into node-level

  • **Network-Centric Community**
    • Partition the whole network into several disjoint sets

  • **Hierarchy-Centric Community**
    • Construct a hierarchical structure of communities
Community Detection

- Node-Centric
- Group-Centric
- Network-Centric
- Hierarchy-Centric
Node-Centric Community Detection

• Nodes satisfy different properties
  • Complete Mutuality
    • cliques
  • Reachability of members
    • k-clique, k-clan, k-club
  • Nodal degrees
    • k-plex, k-core
  • Relative frequency of Within-Outside Ties
    • LS sets, Lambda sets

• Commonly used in traditional social network analysis
Complete Mutuality: Cliques

- **Clique**: a maximum complete subgraph in which all nodes are adjacent to each other.

Nodes 5, 6, 7 and 8 form a clique.

- NP-hard to find the maximum clique in a network.
- Straightforward implementation to find cliques is very expensive in time complexity.
Finding the Maximum Clique

- In a clique of size k, each node maintains degree \( \geq k-1 \)
  - Nodes with degree \(< k-1\) will not be included in the maximum clique
- Recursively apply the following pruning procedure
  - Sample a sub-network from the given network, and find a clique in the sub-network, say, by a greedy approach
  - Suppose the clique above is size k, in order to find out a larger clique, all nodes with degree \(\leq k-1\) should be removed.
- Repeat until the network is small enough
- Many nodes will be pruned as social media networks follow a power law distribution for node degrees
Maximum Clique Example

• Suppose we sample a sub-network with nodes {1-9} and find a clique \{1, 2, 3\} of size 3

• In order to find a clique >3, remove all nodes with degree \( \leq 3 - 1 = 2 \)
  • Remove nodes 2 and 9
  • Remove nodes 1 and 3
  • Remove node 4
Clique Percolation Method (CPM)

• Clique is a very strict definition, unstable
• Normally use cliques as a core or a seed to find larger communities

• CPM is such a method to find overlapping communities
  • Input
    • A parameter \( k \), and a network
  • Procedure
    • Find out all cliques of size \( k \) in a given network
    • Construct a clique graph. Two cliques are adjacent if they share \( k-1 \) nodes
    • Each connected components in the clique graph form a community
CPM Example

 Cliques of size 3: 
 {1, 2, 3}, {1, 3, 4}, {4, 5, 6}, 
 {5, 6, 7}, {5, 6, 8}, {5, 7, 8}, 
 {6, 7, 8}

 Communities:
 {1, 2, 3, 4} 
 {4, 5, 6, 7, 8}
Geodesic

• Reachability is calibrated by the Geodesic distance

• **Geodesic**: a shortest path between two nodes (12 and 6)
  • Two paths: 12-4-1-2-5-6, 12-10-6
  • 12-10-6 is a geodesic

• **Geodesic distance**: #hops in geodesic between two nodes
  • e.g., $d(12, 6) = 2$, $d(3, 11)=5$

• **Diameter**: the maximal geodesic distance for any 2 nodes in a network
  • #hops of the longest shortest path

Diameter = 5
Reachability: k-clique, k-club

- Any node in a group should be reachable in k hops
- **k-clique**: a maximal subgraph in which the largest geodesic distance between any nodes <= k
  - A k-clique can have diameter larger than k within the subgraph
    - e.g., 2-clique \{12, 4, 10, 1, 6\}
- **k-club**: a substructure of diameter <= k
  - e.g., \{1, 2, 5, 6, 8, 9\}, \{12, 4, 10, 1\} are 2-clubs
Nodal Degrees: k-core, k-plex

- Each node should have a certain number of connections to nodes within the group
  - **k-core**: a substructure that each node connects to at least $k$ members within the group
  - **k-plex**: for a group with $n_s$ nodes, each node should be adjacent no fewer than $n_s - k$ in the group

- The definitions are complementary
  - A k-core is a $(n_s - k)$-plex
Within-Outside Ties: LS sets

- **LS sets**: Any of its proper subsets has more ties to other nodes in the group than outside the group

- Too strict, not reasonable for network analysis
Recap of Node-Centric Communities

• Each node has to satisfy certain properties
  • Complete mutuality
  • Reachability
  • Nodal degrees
  • Within-Outside Ties

• Limitations:
  • Too strict, but can be used as the core of a community
  • Not scalable, commonly used in network analysis with small-size network
  • Sometimes not consistent with property of large-scale networks
    • e.g., nodal degrees for scale-free networks
Community Detection

- Node-Centric
- Group-Centric
- Hierarchy-Centric
- Network-Centric
Group-Centric Community Detection

• Consider the connections within a group as whole,
• Some nodes may have low connectivity
• A subgraph with $V_s$ nodes and $E_s$ edges is a $\gamma$-dense quasi-clique if

$$\frac{E_s}{V_s(V_s - 1)/2} \geq \gamma$$

• Recursive pruning:
  • Sample a subgraph, find a maximal $\gamma$-dense quasi-clique
    • the resultant size = $k$
  • Remove the nodes that
    • whose degree < $k\gamma$
    • all their neighbors with degree < $k\gamma$
Community Detection

- Node-Centric
- Group-Centric
- Hierarchy-Centric
- Network-Centric
Network-Centric Community Detection

• To form a group, we need to consider the connections of the nodes globally.

• Goal: partition the network into disjoint sets

• Groups based on
  • Node Similarity
  • Latent Space Model
  • Block Model Approximation
  • Cut Minimization
  • Modularity Maximization
Node Similarity

- Node similarity is defined by how similar their interaction patterns are.
- Two nodes are **structurally equivalent** if they connect to the same set of actors.
  - e.g., nodes 8 and 9 are structurally equivalent.
- Groups are defined over equivalent nodes.
  - Too strict.
  - Rarely occur in a large-scale.
  - Relaxed equivalence class is difficult to compute.
- In practice, use **vector similarity**.
  - e.g., cosine similarity, Jaccard similarity.
**Vector Similarity**

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>

**Cosine Similarity:**

$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|}.$$  

$$\text{sim}(5,8) = \frac{1}{\sqrt{2} \times \sqrt{3}} = \frac{1}{\sqrt{6}}$$

**Jaccard Similarity:**

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}.$$  

$$J(5,8) = \frac{|\{6\}|}{|\{1,2,6,13\}|} = \frac{1}{4}$$
Clustering based on Node Similarity

• For practical use with huge networks:
  • Consider the connections as features
  • Use Cosine or Jaccard similarity to compute vertex similarity
  • Apply classical k-means clustering Algorithm

• K-means Clustering Algorithm
  • Each cluster is associated with a centroid (center point)
  • Each node is assigned to the cluster with the closest centroid

Algorithm 1 Basic K-means Algorithm.

1: Select $K$ points as the initial centroids.
2: **repeat**
3: Form $K$ clusters by assigning all points to the closest centroid.
4: Recompute the centroid of each cluster.
5: **until** The centroids don’t change
Cut-Minimization

- Between-group interactions should be infrequent
- **Cut**: number of edges between two sets of nodes
- **Objective**: minir
  \[
  \text{cut}(C_1, C_2, \cdots, C_k) = \sum_{i=1}^{k} \text{cut}(C_i, \overline{C_i})
  \]
  - Limitations: often find communities of only one node
  - Need to consider the group size
- Two commonly-used variants:
  - **Cut = 1**: Number of nodes in a community
  - **Cut = 2**: Number of within-group Interactions
  - **Ratio-cut**: Number of within-group Interactions
  \[
  \text{Ratio-cut}(C_1, C_2, \cdots, C_k) = \sum_{i=1}^{k} \frac{\text{cut}(C_i, \overline{C_i})}{|V_i|}
  \]
  - **Normalized-cut**: Number of within-group Interactions
  \[
  \text{Normalized-cut}(C_1, C_2, \cdots, C_k) = \sum_{i=1}^{k} \frac{\text{cut}(C_i, \overline{C_i})}{\text{vol}(V_i)}
  \]
Modularity Maximization

- **Modularity** measures the group interactions compared with the expected random connections in the group.
- In a network with \( m \) edges, for two nodes with degree \( d_i \) and \( d_j \), expected random connections between them are
  \[
  \frac{d_i d_j}{2m}
  \]

- The interaction utility in a group:
  \[
  \sum_{i \in C, j \in C} A_{ij} - \frac{d_i d_j}{2m}
  \]

- To partition the group into multiple groups, we maximize
  \[
  \frac{1}{2m} \sum_{C} \sum_{i \in C, j \in C} A_{ij} - \frac{d_i d_j}{2m}
  \]

Expected Number of edges between 6 and 9 is
\[
\frac{5 \times 3}{2 \times 17} = \frac{15}{34}
\]
Properties of Modularity

• Properties of modularity:
  • Between (-1, 1)
  • Modularity = 0 if all nodes are clustered into one group
  • Can automatically determine optimal number of clusters

• Resolution limit of modularity
  • Modularity maximization might return a community consisting multiple small modules
A Unified View for Community Partition

- Latent space models, block models, spectral clustering, and modularity maximization can be unified as

Utility Matrix $M = \begin{cases} 
\text{modified proximity matrix } \tilde{P} & \text{if latent space models} \\
\text{adjacency matrix } A & \text{if block models} \\
\text{graph Laplacian } \tilde{L} & \text{if spectral clustering} \\
\text{modularity maximization } B & \text{if modularity maximization} 
\end{cases}$

Reference: http://www.cse.ust.hk/~weikep/notes/Script_community_detection.m
Recap of Network-Centric Community

• **Goal**: Partition network nodes into several disjoint sets

• **Limitation**: Require the user to specify the number of communities beforehand
Hierarchy-Centric Community Detection

- Community Detection
  - Node-Centric
  - Group-Centric
  - Network-Centric
  - Hierarchy-Centric
Hierarchy-Centric Community Detection

• **Goal:** Build a hierarchical structure of communities based on network topology

• Facilitate the analysis at different resolutions

• Representative Approaches:
  • Divisive Hierarchical Clustering
  • Agglomerative Hierarchical Clustering
Link Mining
What is Link Mining?

• Traditional machine learning and data mining approaches assume:
  • A random sample of homogeneous objects from single relation
• Real world data sets:
  • Multi-relational, heterogeneous and semi-structured
• Link Mining
  • Newly emerging research area at the intersection of research in social network and link analysis, hypertext and web mining, graph mining, relational learning and inductive logic programming
What Is a Link in Link Mining?

• Link: relationship among data
• Two kinds of linked networks
  • homogeneous vs. heterogeneous
• Homogeneous networks
  • Single object type and single link type
  • Single model social networks (e.g., friends)
  • WWW: a collection of linked Web pages
• Heterogeneous networks
  • Multiple object and link types
  • Medical network: patients, doctors, disease, contacts, treatments
  • Bibliographic network: publications, authors, venues
Link mining

• Availability of rich data on link structure between objects
• Link Mining - new emerging field encompassing a range of tasks including descriptive and predictive modeling (Getoor, 2003)
• Extending classical data mining tasks
  – Link-based classification – predict an object’s category based not only on its attributes but also the links it participates in
  – Link-based clustering – techniques grouping objects (or linked objects)
• Special cases of link-based classification/clustering
  – Identifying link type
  – Predicting link strength
  – Link cardinality
• Getoor et al (2002)
  – Two mechanisms to represent probabilistic distributions over link structures
  – Apply resulting model to predict link structure
Link Prediction

• Predict whether a link exists between two entities, based on attributes and other observed links

• Applications
  • Web: predict if there will be a link between two pages
  • Citation: predicting if a paper will cite another paper
  • Epidemics: predicting who a patient’s contacts are

• Methods
  • Often viewed as a binary classification problem
  • Local conditional probability model, based on structural and attribute features
  • Difficulty: sparseness of existing links
  • Collective prediction, e.g., Markov random field model
Link Prediction

• **Different versions**
  – Given a social network at time $t_i$ predict the social link between actors at time $t_{i+1}$
  – Given a social network with an *incomplete* set of social links between a *complete* set of actors, predict the unobserved social links
  – Given information about actors, predict the social link between them (this is quite similar to social network extraction)

  ![Diagram](image)

  - Time $t$
  - Time $(t+1)$
  - Incomplete Network

• **Classical approach for link prediction is to fit the social network on a model and then use it for link prediction**

• **Link Mining - encompassing a range of tasks including descriptive and predictive modelling (Getoor, 2003)**
Link Prediction

• Predictive powers of the various proximity features for predicting links between authors in the future (Liben-Nowell and Kleinberg, 2003)
  – Link prediction as a means to gauge the usefulness of a model
  – Proximity Features: Common Neighbors, Katz, Jaccard, etc
  – No single predictor consistently outperforms the others
    • However all perform better than random

• Link Prediction using supervised learning (Hasan et al, 2006)
  – Citation Network (BIOBASE, DBLP)
  – Use machine learning algorithms to predict future co-authorship (decision tree, k-NN, multilayer perceptron, SVM, RBF network)
  – Identify a group of features that are most helpful in prediction
  – Best Predictor Features: Keyword Match count, Sum of neighbors, Sum of Papers, Shortest Distance
Link Prediction

- Prediction of Link Attachments by Estimating Probabilities of Information Propagation (Saito et al 2007)
- Problem: Given a network at time $t$, the goal is to predict $k$ potential links that are most likely to be converted to real links after a certain period of time.
- A ranking method: Top $k$ links are predicted to be the real links.
- Pick two nodes $v$ and $w$ such that edge $(v,w)$ does not exist and $d(v,w) = 2$
- An edge is created between $v$ and the adjacent nodes of $w$ if information propagation between the two is successful.
- In the dataset only a small fraction (0.0002) of the potential links are converted to real links. The proposed method outperformed all the other comparison methods.
Link Prediction using Collaborative Filtering

- Find the background model that can generate the link data
<table>
<thead>
<tr>
<th>User</th>
<th>Item 1</th>
<th>Item 2</th>
<th>Item 3</th>
<th>Item 4</th>
<th>Item 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>User 1</td>
<td>1</td>
<td>?</td>
<td>2</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>User 2</td>
<td>?</td>
<td>5</td>
<td>7</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>User 3</td>
<td>4</td>
<td>7</td>
<td>4</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>User 4</td>
<td>7</td>
<td>1</td>
<td>7</td>
<td>3</td>
<td>8</td>
</tr>
<tr>
<td>User 5</td>
<td>7</td>
<td>4</td>
<td>6</td>
<td>?</td>
<td></td>
</tr>
<tr>
<td>User 6</td>
<td>3</td>
<td>8</td>
<td>3</td>
<td>7</td>
<td></td>
</tr>
</tbody>
</table>
Challenges in Link Prediction

• Data!!!

• Cold Start Problem

• Sparsity Problem
Link Prediction using Collaborative Filtering

• Memory-based Approach
  • User-base approach [Twitter]
  • item-base approach [Amazon & Youtube]

• Model-based Approach
  • Latent Factor Model [Google News]

• Hybrid Approach
Memory-based Approach

• Few modeling assumptions
• Few tuning parameters to learn
• Easy to explain to users

• Dear Amazon.com Customer, We've noticed that customers who have purchased or rated *How Does the Show Go On: An Introduction to the Theater* by Thomas Schumacher have also purchased *Princess Protection Program #1: A Royal Makeover* (*Disney Early Readers*).
Algorithms: User-Based Algorithms (Breese et al, UAI98)

• $v_{i,j}$ = vote of user $i$ on item $j$
• $I_i$ = items for which user $i$ has voted
• Mean vote for $i$

$$
\bar{v}_i = \frac{1}{|I_i|} \sum_{j \in I_i} v_{i,j}
$$

• Predicted vote for “active user” $a$ is weighted sum

$$
p_{a,j} = \bar{v}_a + \kappa \sum_{i=1}^{n} w(a, i) (v_{i,j} - \bar{v}_i)
$$

(normalizer) weights of $n$ similar users
Algorithm: Amazon’s Method

• Item-based Approach
  • Similar with user-based approach but is on the item side
### Item-based CF Example: infer (user 1, item 3)

<table>
<thead>
<tr>
<th></th>
<th>Item 1</th>
<th>Item 2</th>
<th>Item 3</th>
<th>Item 4</th>
<th>Item 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>User 1</td>
<td>8</td>
<td>1</td>
<td>?</td>
<td>2</td>
<td>7</td>
</tr>
<tr>
<td>User 2</td>
<td>2</td>
<td>?</td>
<td>5</td>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td>User 3</td>
<td>5</td>
<td>4</td>
<td>7</td>
<td>4</td>
<td>7</td>
</tr>
<tr>
<td>User 4</td>
<td>7</td>
<td>1</td>
<td>7</td>
<td>3</td>
<td>8</td>
</tr>
<tr>
<td>User 5</td>
<td>1</td>
<td>7</td>
<td>4</td>
<td>6</td>
<td>?</td>
</tr>
<tr>
<td>User 6</td>
<td>8</td>
<td>3</td>
<td>8</td>
<td>3</td>
<td>7</td>
</tr>
</tbody>
</table>
How to Calculate Similarity (Item 3 and Item 5)?

<table>
<thead>
<tr>
<th></th>
<th>Item 1</th>
<th>Item 2</th>
<th>Item 3</th>
<th>Item 4</th>
<th>Item 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>User 1</td>
<td>8</td>
<td>1</td>
<td>?</td>
<td>2</td>
<td>7</td>
</tr>
<tr>
<td>User 2</td>
<td>2</td>
<td>?</td>
<td>5</td>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td>User 3</td>
<td>5</td>
<td>4</td>
<td>7</td>
<td>4</td>
<td>7</td>
</tr>
<tr>
<td>User 4</td>
<td>7</td>
<td>1</td>
<td>7</td>
<td>3</td>
<td>8</td>
</tr>
<tr>
<td>User 5</td>
<td>1</td>
<td>7</td>
<td>4</td>
<td>6</td>
<td>?</td>
</tr>
<tr>
<td>User 6</td>
<td>8</td>
<td>3</td>
<td>8</td>
<td>3</td>
<td>7</td>
</tr>
</tbody>
</table>
### Similarity between Items

#### Table

<table>
<thead>
<tr>
<th>Item 3</th>
<th>Item 4</th>
<th>Item 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>?</td>
<td>2</td>
<td>7</td>
</tr>
<tr>
<td>5</td>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td>7</td>
<td>4</td>
<td>7</td>
</tr>
<tr>
<td>7</td>
<td>3</td>
<td>8</td>
</tr>
<tr>
<td>4</td>
<td>6</td>
<td>?</td>
</tr>
<tr>
<td>8</td>
<td>3</td>
<td>7</td>
</tr>
</tbody>
</table>

- How similar are items 3 and 5?
- How to calculate their similarity?
Similarity between items

<table>
<thead>
<tr>
<th>Item 3</th>
<th>Item 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>?</td>
<td>7</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>4</td>
<td>?</td>
</tr>
<tr>
<td>8</td>
<td>7</td>
</tr>
</tbody>
</table>

- Only consider users who have rated both items
- For each user: Calculate difference in ratings for the two items
- Take the average of this difference over the users

\[
sim(\text{item 3, item 5}) = \cosine\left( (5, 7, 7), (5, 7, 8) \right) \\
= (5*5 + 7*7 + 7*8)/(\sqrt{5^2+7^2+7^2} \ast \sqrt{5^2+7^2+8^2})
\]

- Can also use Pearson Correlation Coefficients as in user-based approaches
Prediction: Calculating ranking $r(user_1, item_3)$

$$r(user_1, item_3) = \alpha \times \{ r(user_1, item_1) \times \text{sim}(item_1, item_3) + r(user_1, item_2) \times \text{sim}(item_2, item_3) + r(user_1, item_4) \times \text{sim}(item_4, item_3) + r(user_1, item_5) \times \text{sim}(item_5, item_3) \}$$

Where $\alpha$ is a normalization factor, which is $1/\text{the sum of all sim}(item_i, item_3)$. 
Where I got some of these slides from.

• (parts of these slides are from EBDT Summer School) Francesco Bonchi Yahoo! Research Barcelona

• Parts of these slides are from: Jiawei Han & Michael Kambers Data Mining: Concepts and Techniques

• Parts of these slides are also from: Jaideep Srivastava, Muhammad A. Ahmad, Nishith Pathak, David Kuo-Wei Hsu. Data Mining Based Social Network Analysis from Online Behaviour. SIAM Data Mining
References
References

3. Ahmad, Muhammad A., Teredesai, Ankur., Modeling Proliferation of Ideas in Online Social Networks, Proceedings of the 5th Australasian Data Mining Conference, November 29-30 2006, Held in conjunction with the 19th Australian Joint Conference on Artificial Intelligence, Sydney, AUS.
References


References


References


46. David Kempe, Jon M. Kleinberg, Éva Tardos: Maximizing the spread of influence through a social network. KDD 2003: 137-146


48. Young Ae Kim, Jaideep Srivastava: Impact of social influence in e-commerce decision making. ICEC 2007: 293-302


59. Jure Leskovec, Lada Adamic, and Bernardo Huberman. The Dynamics of Viral Marketing. EC'06.


References


References


References


96. E. Spertus, Mehran Sahami, Orkut Buyukkokten: Evaluating similarity measures: a large-scale study in the orkut social network. KDD 2005: 678-684


References


Search in Social Networks
Expertise in Social Networks

- Expertise oriented search using social networks (Jing Zhang 2007, Juan-Zi Lee 2007)
- A social network is constructed the co-authorship between authors
- Expert Identification
  - First compute relevancy based on documents associated with the author for a given topic.
  - Secondly propagate the topic relevancy of the researcher to his/her neighbors.
  - Thus the expertise depends upon authored documents and the expertise of one’s neighbors.
  - Alternatively, compute expertise and then get the experts relevant to the query and then construct the social network and then propagate expertise.

![An example of an academic researcher network](image)

In general, the expert score \( s(v_j)^{t+1} \) is computed from \( s(v_j)^t \) as follows (normalization is omitted for clarity):

\[
s(v_j)^{t+1} = s(v_j)^t + \sum_{v_j \in U} \sum_{e \in R_{ji}} w((v_j, v_i), e) s(v_j)^t
\]

(1)

where \( w((v_j, v_i), e) \) represents the propagation coefficient and \( e \in R_{ji} \) is one kind of relationship from the person \( v_j \) to \( v_i \); \( U \) stands for all neighboring nodes to \( v_i \) in the graph and \( R_{ji} \) stands for all relationships from the person \( v_j \) to \( v_i \).
Search in Social Networks

- Yu and Singh (2003)
  - Each actor has a vector over all terms and every actor stores the vectors and immediate neighborhoods of his/her neighbors.
  - Individual vector entries indicate actor’s familiarity/knowledge about the various terms.
  - Each neighbor is assigned a relevance score.
  - The score is a weighted linear combination of the similarity between query and term vectors (cosine similarity based measure) and the sociability of that neighbor.
  - Sociability is a measure of that neighbor knowing other people who might know the answer.
  - The expert and sociability ratings maintained by a user are updated based on answers provided by various users in the network.
Search and Expert Identification

- Setting: A decentralized knowledge market in the form of a social network
- Goal: Identify experts and route queries in the network
- Solution: An ant colony optimization technique that keeps track of the history of past queries.

- Advantages:
  - No need to keep track of ‘topics’ as topics can evolve.
  - Takes into account the dynamic nature of the network.
  - Track the changes in expertise of nodes over time.

(Ahmad and Srivastava 2008) Related Work: (Yu and Singh, 2003)
DIFSoN

- The DIFSoN (Leung et al. 2014) uses
  - (i) a prefix-tree based data structure called Influential Friend tree (IF-tree) to effectively capture the social network data
  - (ii) a mining routine to efficiently discover the set of influential friend groups from the IF-tree.

<table>
<thead>
<tr>
<th>Table 1</th>
<th>A collection LC of friend lists</th>
</tr>
</thead>
<tbody>
<tr>
<td>List ID</td>
<td>Friend list (L)</td>
</tr>
<tr>
<td>L₁</td>
<td>{Ana, Beto, Carlos, Eva}</td>
</tr>
<tr>
<td>L₂</td>
<td>{Ana, Beto, Carlos}</td>
</tr>
<tr>
<td>L₃</td>
<td>{Beto, Eva, Fabio}</td>
</tr>
<tr>
<td>L₄</td>
<td>{Ana, Beto, Davi}</td>
</tr>
<tr>
<td>L₅</td>
<td>{Ana, Beto, Carlos, Eva}</td>
</tr>
<tr>
<td>L₆</td>
<td>{Beto, Eva, Fabio}</td>
</tr>
<tr>
<td>L₇</td>
<td>{Ana, Beto, Carlos, Eva}</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 2</th>
<th>Prominence of friends in a social network</th>
</tr>
</thead>
<tbody>
<tr>
<td>Friend (fᵢ)</td>
<td>Prominence (Prom(fᵢ))</td>
</tr>
<tr>
<td>Ana</td>
<td>0.60</td>
</tr>
<tr>
<td>Beto</td>
<td>0.50</td>
</tr>
<tr>
<td>Carlos</td>
<td>0.40</td>
</tr>
<tr>
<td>Davi</td>
<td>0.70</td>
</tr>
<tr>
<td>Eva</td>
<td>0.42</td>
</tr>
<tr>
<td>Fabio</td>
<td>0.57</td>
</tr>
<tr>
<td>Gil</td>
<td>0.11</td>
</tr>
</tbody>
</table>
DIFSoN

• The prominence $Prom(G)$ of a group $G$ measures the average of all prominence values for all friends in the group:

$$Prom(G) = \frac{\sum_{i=1}^{size(G)} Prom(f_i)}{size(G)}.$$

• The influence $Inf(G, LC)$ of a group $G$ in a collection $LC$ of friend lists measures an aggregated prominence degree of $G$ in all friend lists in $LC$. It is defined as the product of the prominence value of $G$ and its appearance frequency in $LC$ [i.e., $Freq(G, LC)$]:

$$Inf(G, LC) = Prom(G) \times Freq(G, LC).$$
DIFSoN

• Given a user-specified minimum social network influence threshold \( \text{minInf} \), a group \( G \) is considered influential in a collection \( LC \) of friend lists if its influence value is at least \( \text{minInf} \):

\[
\text{Inf}(G, LC) \geq \text{minInf}.
\]
Influential Friend tree (IF-tree)

(a) After inserting $L_1$

(b) After inserting $L_2$

(c) After inserting $L_3$

(d) After inserting all lists

(e) After deleting uninfluential friends

Fig. 1 The IF-tree construction
Influential Friend tree (IF-tree)

(a) \{Ana\}-projected tree, which is the same as \{Ana\}-conditional tree

(b) \{Ana, Beto\}-conditional tree

(c) \{Ana, Beto, Eva\}-conditional tree

(d) \{Ana, Eva\}-conditional tree

(e) \{Beto\}-conditional tree

(f) \{Beto, Eva\}-projected tree

Fig. 2 The IF-tree mining with PromGMax
Trust in Social Networks
Trust in Social Networks

- **Trust propagation**: An approach for inferring trust values in a network
  - A user trusts some of his friends, his/her friends trust their friends and so on...
  - Given trust and/or distrust values between a handful of pairs of users, can one predict unknown trust/distrust values between any two users


- TrustMail
  - Consider research groups X and Y headed by two professors such that each professor knows the students in their respective group
  - If a student from group X sends a mail to the professor of group Y then how will the student be rated?
  - Use the rating of professor from group X who is in professor Y's list of trusted list and propagate the rating

- Example of a real life trust model – [www.ebay.com](http://www.ebay.com)
Trust in Social Networks

- **TidalTrust Algorithm** (Golbeck, 2005)
  - A source is more likely to believe the trust ratings, regarding a third person (sink), from a *close* and *highly trusted* actor
  - Using BFS all paths with the minimum length from source to sink are determined
  - Trust rating for a path is the minimum trust rating along that path
  - Use weighted average of trust ratings only from those paths on which source trusts its neighbour > max {trust score of all paths}

- **Propagation of Trust and Distrust in Networks** (Guha et al, 2004)
  - Propose a framework of trust propagation schemes
  - Modelled via a matrix of Beliefs $B = T$ (Trust matrix) or $B = T-D$ (Trust – Distrust)
  - Applications of atomic propagations are used to propagate trust values
    - E.g. Trust is transitive - $B*B$, co-citation - $B*B^T B$
  - Various schemes for chaining atomic propagations
  - **Goal**: Produce a final matrix $F$ from which one can read off the computed trust or distrust of any two users
Application of Data Mining Based Social Network Analysis Techniques
Applications (Outline)

• Organization Theory
• Semantic Web
• Viral Marketing
• Social Influence and E-Commerce
• Social Computing
• Criminal Network Analysis
• Newsgroup Message Classification
• Social Recommendation Systems
• Terrorism and Crime Related Weblog Social Network
Organization Theory

- Krackhardt and Hanson (1993)
  - Informal (social) networks present in an enterprise are different from formal networks
  - Different patterns exist in such networks like imploded relationships, irregular communication patterns, fragile structures, holes in network and bow ties

  - Survey as well as study the impact of informal networks on an enterprise

(Source: Krackhardt and Hanson, 1993)
Extracting Co-appearance Networks among Organizations

- Extracting Inter-Firm Networks from WWW (Jin et al., 2007)

Results form a search engine can be estimated in a more robust way (Matsuo et al., 2007)

Query about Relation (Link analysis)
"Matsushita AND JustSystem"

Too many pages

Content analysis

Top-ranked pages are about lawsuit

Query about Relation and Relation keyword
"Matsushita AND JustSystem AND lawsuit"

Figure 2. System flow to extract a firm network.
Semantic Web Community

• Ding et al (2005)
  – Semantic web enables explicit, online representation of social information while social networks provide a new paradigm for knowledge management e.g. Friend-of-a-friend (FOAF) project (http://www.foaf-project.org)
  – Applied SNA techniques to study this FOAF data (DS-FOAF)

Preliminary analysis of DS-FOAF data (Ding et al, 2005)

Figure 3: Cumulative distribution of in-degree and out-degree

Degree distribution

Connected components

Figure 4: We found that connected components in FOAF networks followed a few simple patterns.

Trust across multiple sources (Ding et al, 2005)
Semantic Web and SNA

- The friend of a friend (FOAF) project has enabled collection of machine readable data on online social interactions between individuals. [http://www.foaf-project.org](http://www.foaf-project.org)

The Sun never sets under the Semantic Web: the network of semantic web researchers across globe (Mika, 2005)

Snapshot of clusters ([http://flink.semanticweb.org/](http://flink.semanticweb.org/))
Viral Marketing

• Domingos(2005), Domingos and Richardson (2001, 2002)
  – *Network value* of a customer is the expected profit from marketing a product to a customer, accounting for the customer’s influence on the buying decisions of other customers
  – Propose a greedy strategy for identifying customers with maximum network value

• Kempe et al (2003)
  • For a general class of cascading models, the problem of identifying customers with maximum network value is NP-hard
  • A greedy strategy provides a solution within 63% of the optimal

(Source: Leskovec et al, 2006)
Criminal Network Analysis

- Example (Qin et al, 2005)
  - Information collected on social relations between members of Global Salafi Jihad (GSJ) network from multiple sources (e.g. reports of court proceedings)
  - Applied social network analysis as well as Web structural mining to this network
  - Authority derivation graph (ADG) captures (directed) authority in the criminal network

<table>
<thead>
<tr>
<th>Ranking</th>
<th>Leader</th>
<th>Gatekeeper</th>
<th>Outlier</th>
</tr>
</thead>
<tbody>
<tr>
<td>Central Member</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>Zawahiri</td>
<td>bin Laden</td>
<td>Khallaf</td>
</tr>
<tr>
<td>2</td>
<td>Makkawi</td>
<td>Zawahiri</td>
<td>Stin Laden</td>
</tr>
<tr>
<td>3</td>
<td>Ibn al-Shaykh</td>
<td>Khadr</td>
<td>Hashk</td>
</tr>
<tr>
<td>4</td>
<td>bin Laden</td>
<td>Sirri</td>
<td>M Afef</td>
</tr>
<tr>
<td>5</td>
<td>Alaa ibn Aayad</td>
<td>Zawahiri</td>
<td>Sheikh Omar</td>
</tr>
</tbody>
</table>

Core Arab
- Khalid | Harithi | El Barahi
- Shaib | Nasiri | Khaddaf
- Jarrahi | Khalid | Baghdad
- AlTa | Johani | Dabab
- AlShahar | Zaiblu | Mehdi

Maghrib Arab
- Haabali | Baayr | Biskr
- Baayr | Haabali | Futha
- Muhib | Guban | Nabardin
- Aybal | Mubajir | Yunos2
- Azahari | Sefeno | Mohib

Southeast Asian
- Doha | Yarski | Mijjati
- Benyaclu2 | Zauri | Perlia
- Fath | Chah | Mahdjoub
- Chah | Darvul | Znline
- Benyaclu1 | Maanoufi | Ziyad

Terrorists with top centrality ranks in each clump

1-hop network of 9/11 attack

ADG of GSJ network
Criminal Network Analysis

- Knowledge gained by applying SNA to criminal network aids law enforcement agencies to fight crime proactively.
- Criminal networks are large, dynamic and characterized by uncertainty.
- Need to integrate information from multiple sources (criminal incidents) to discover regular patterns of structure, operation and information flow (Xu and Chen, 2005).
- Computing SNA measures like centrality is NP-hard.
  - **Approximation techniques** (Carpenter et al 2002)
- Visualization techniques for such criminal networks are needed.

*Figure: Terrorist network of 9/11 hijackers (Krebs, 2001/ Xu and Chen, 2005)*

Example of 1st generation visualization tool.  Example of 2nd generation visualization tool.
Newsgroup Message Classification

- Using SNA to help classify newsgroup messages (Fortuna et. Al, 2007)
  - SVM classifier
  - Rich feature set from “networks”

Networks where users socially interact with others through posting and replying

Networks where similarities between two nodes are determined by authors or contents
Social Recommendation Systems

• Initial approaches
  – Anonymous recommendations: treat individuals preferences as independent of each other
  – Failure to account for influence of individual’s social network on his/her preferences
• Kautz et al (1997)
  – Incorporate information of social networks into recommendation systems
  – Enables more focused and effective search
• McDonald (2003)
  – Analyzes the use of social networks in recommendation systems
  – Highlights the need to balance between purely social match vs. expert match
  – Aggregate social networks may not work best for individuals
  – Apply social network analysis techniques to represent & analyze collaboration in recommender systems
• Lam (2004)
  – SNACK - an automated collaborative system that incorporates social information for recommendations
  – Mitigates the problem of cold-start, i.e. recommending to a user who not yet specified preferences
Conclusion

• Computers have provided the ideal infrastructure for
  – Fostering social interaction
  – Capture it at a very fine granularity
  – Practically no reporting bias

• -+ Fertile research area for data mining research