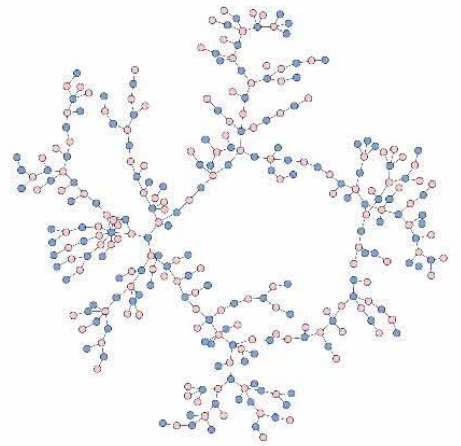
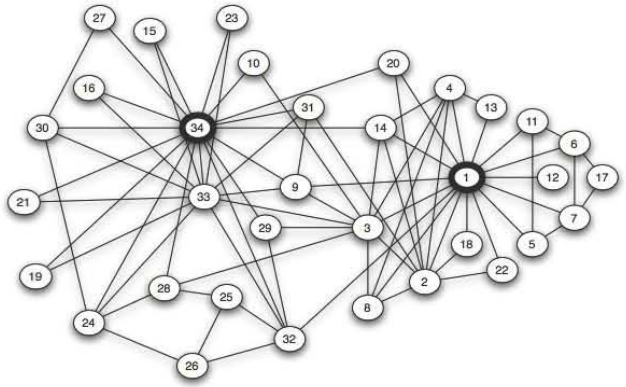
Social Network Mining

Social Network Analysis



High-school dating (Bearman-Moody-Stovel 2004)

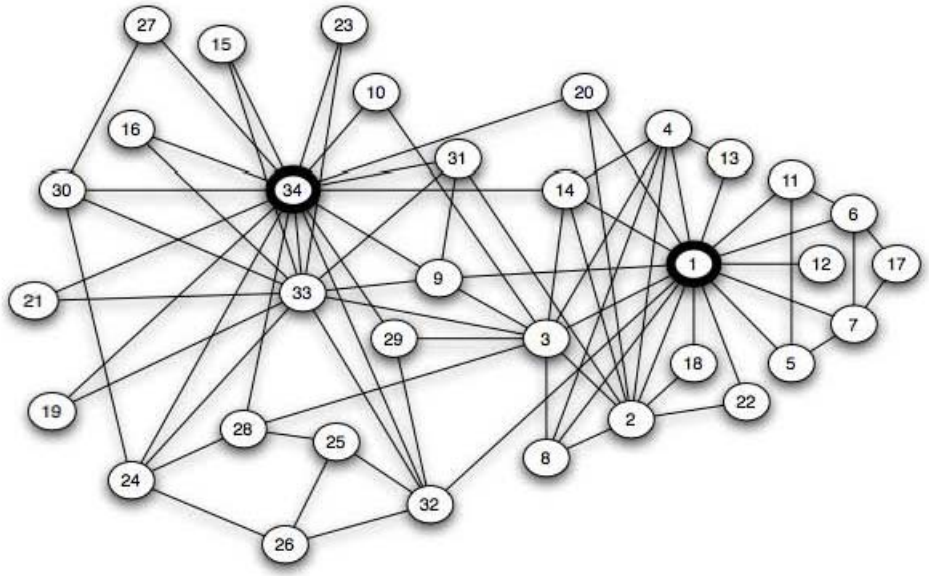


Karate club (Zachary 1977)

Social network data

- Active research area in sociology, social psychology, anthropology for the past half-century.
- Today: Convergence of social and technological networks Computing and info. systems with intrinsic social structure.
- What can the different fields learn from each other?

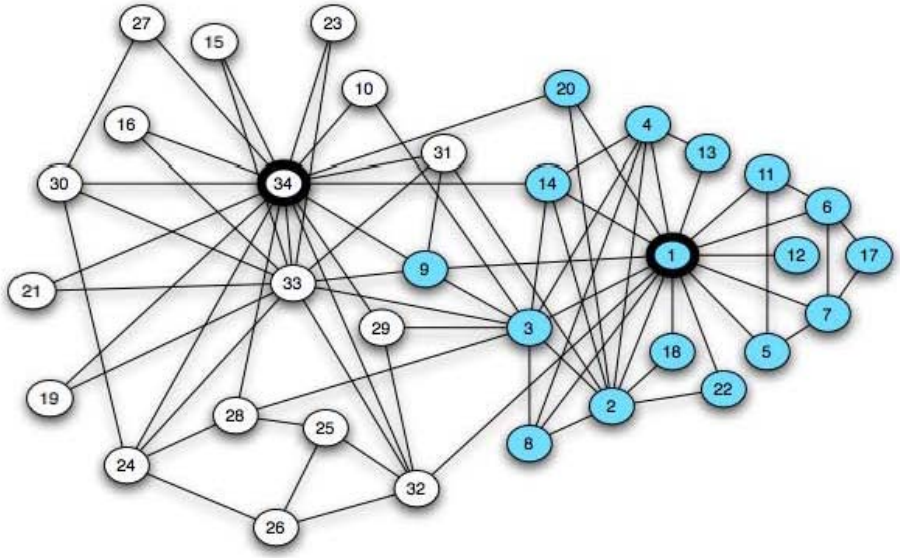
Mining Social Network Data



Mining social networks also has long history in social sciences.

- E.g. Wayne Zachary's Ph.D. work (1970-72): observe social ties and rivalries in a university karate club.

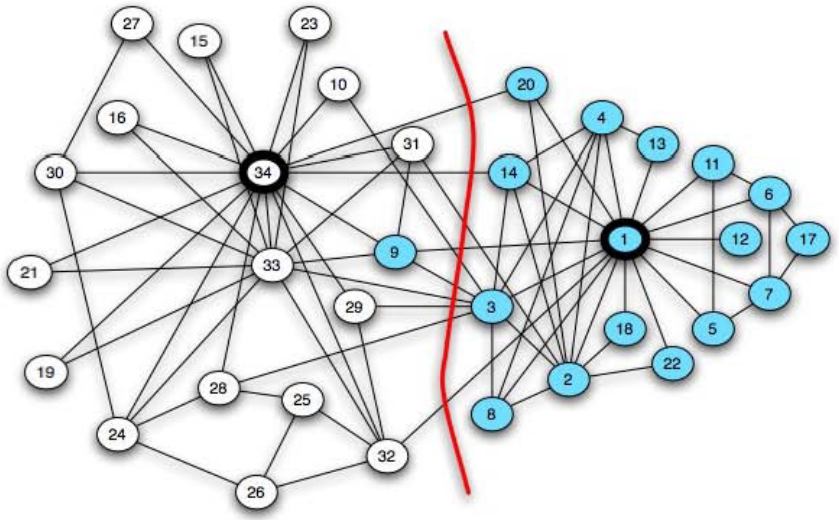
Mining Social Network Data



Mining social networks also has long history in social sciences.

- E.g. Wayne Zachary's Ph.D. work (1970-72): observe social ties and rivalries in a university karate club.
- During his observation, conflicts intensified and group split.

Mining Social Network Data



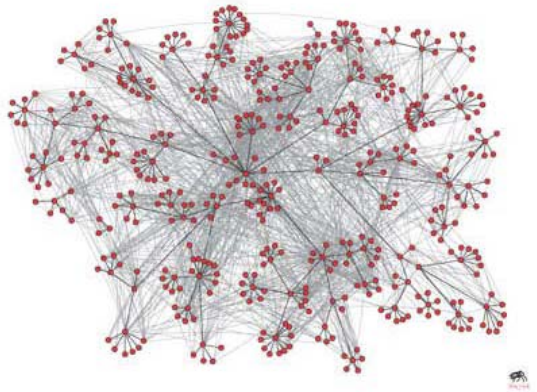
Mining social networks also has long history in social sciences.

- E.g. Wayne Zachary's Ph.D. work (1970-72): observe social ties and rivalries in a university karate club.
- During his observation, conflicts intensified and group split.
- Split could be explained by minimum cut in social network.

A matter of Scale

Social network data spans many orders of magnitude

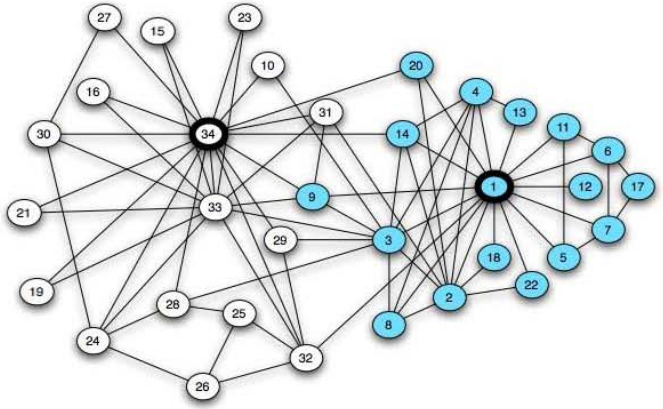
- 436-node network of e-mail exchange over 3 months at a corporate research lab (Adamic-Adar 2003)



- 43,553-node network of e-mail exchange over 2 years at a large university (Kossinets-Watts 2006)
- 4.4-million-node network of declared friendships on blogging community LiveJournal (Liben-Nowell et al. 2005, Backstrom et al. 2006)
- 240-million-node network of all IM communication over one month on Microsoft Instant Messenger (Leskovec-Horvitz'07)

Not just a matter of Scale

- How does massive network data compare to small-scale studies?



Currently, massive network datasets give you both more and less:

- More: can observe global phenomena that are genuine, but literally invisible at smaller scales.
- Less: Don't really know what any one node or link means. Easy to measure things; hard to pose nuanced questions.
- Goal: Find the point where the lines of research converge.

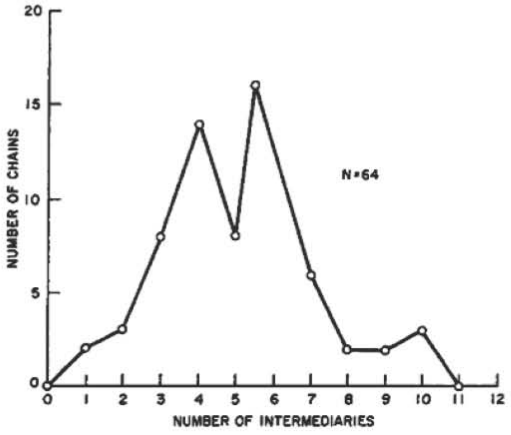
Several core KDD methodologies come into play

- Working with network data that is much messier than just nodes and edges.
- Algorithmic models as a basic vocabulary for expressing complex social-science questions on complex network data.
- Understanding social networks as datasets: privacy implications and other concerns.

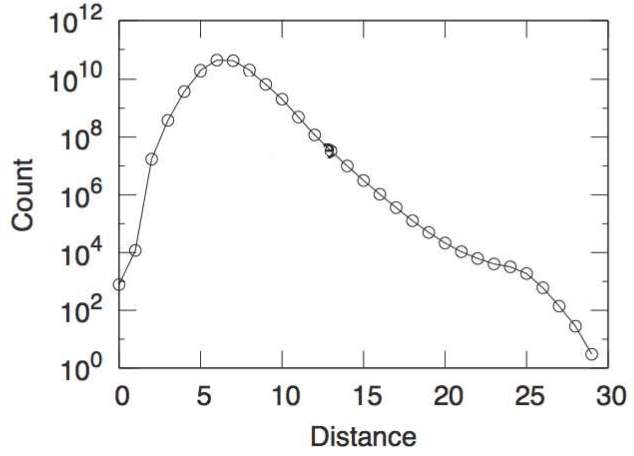
The Small-Worls Phenomenon

Milgram's small-world experiment (1967)

Choose a target in Boston, starters in Nebraska.
A letter begins at each starter, must be passed between personal acquaintances until target is reached.
Six steps on average $\longrightarrow$ six degrees of separation.

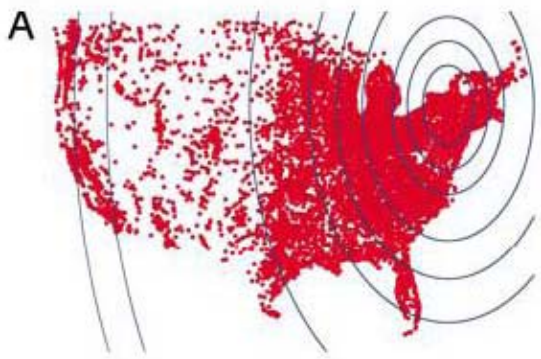


Milgram experiment (Travers-Milgram 1970)



Microsoft IM (Leskovec-Horvitz 2007)

Geographic Data: Live Journal



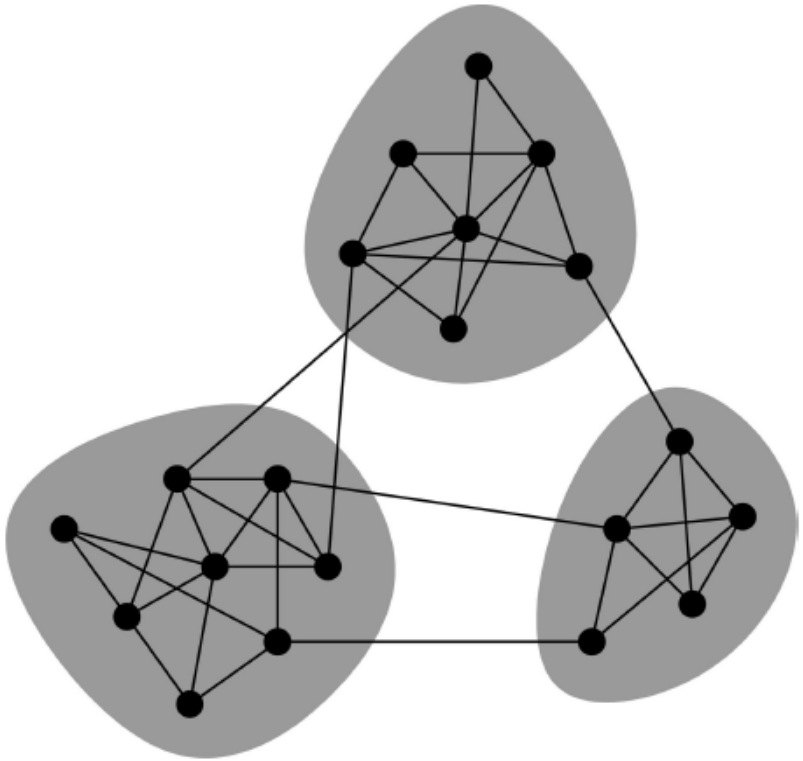
Liben-Nowell, Kumar, Novak, Raghavan, Tomkins (2005) studied LiveJournal, an on-line blogging community with friendship links.

- Large-scale social network with geographical embedding:
  - 500,000 members with U.S. Zip codes, 4 million links.
- Analyzed how friendship probability decreases with distance.
- Difficulty: non-uniform population density makes simple lattice models hard to apply.

Network Communities

- Networks of **tightly connected groups**

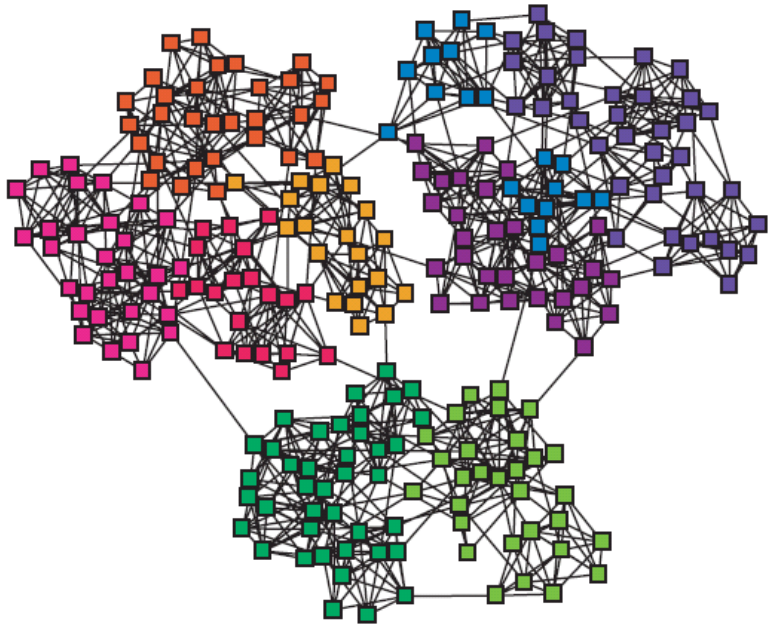
- **Network communities:**
  - Sets of nodes with **lots** of connections **inside** and **few** to **outside** (the rest of the network)



Communities, clusters, groups, modules

Finding Network Communities

- **How to automatically find such densely connected groups of nodes?**
- Ideally such clusters then correspond to real groups
- **For example:**



Communities, clusters, groups, modules

Measures of network centrality

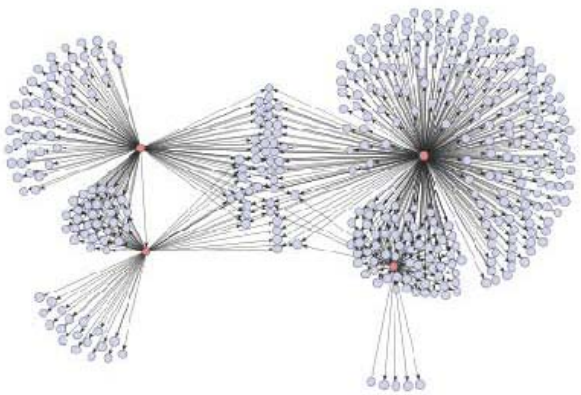
- **Betweenness Centrality:** Measures how many times a node occurs in a shortest path; measure of 'social brokerage power'
  - Most popular measure of centrality
  - Efficient computation is important, best technique is O(mn)

- **Closeness Centrality:** The total graph-theoretic distance of a given node from all other nodes

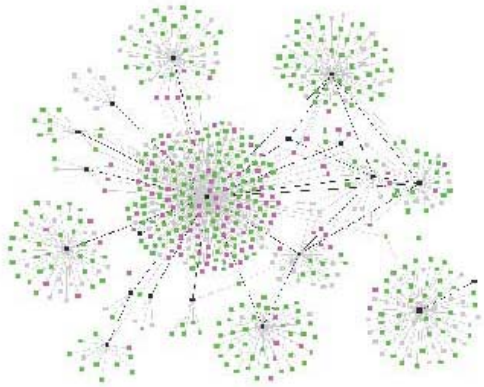
- **Degree centrality:** Degree of a node normalized to the interval {0 .. 1}
  - is in principle identical for egocentric and socio-centric network data

- **Eigenvector centrality:** Score assigned to a node based on the principle that a high scoring neighbour contributes more weight to it
  - Google's PageRank is a special case of this

Diffusion in Social Networks



Book recommendations (Leskovec et al 2006)



Contagion of TB (Andre et al. 2006)

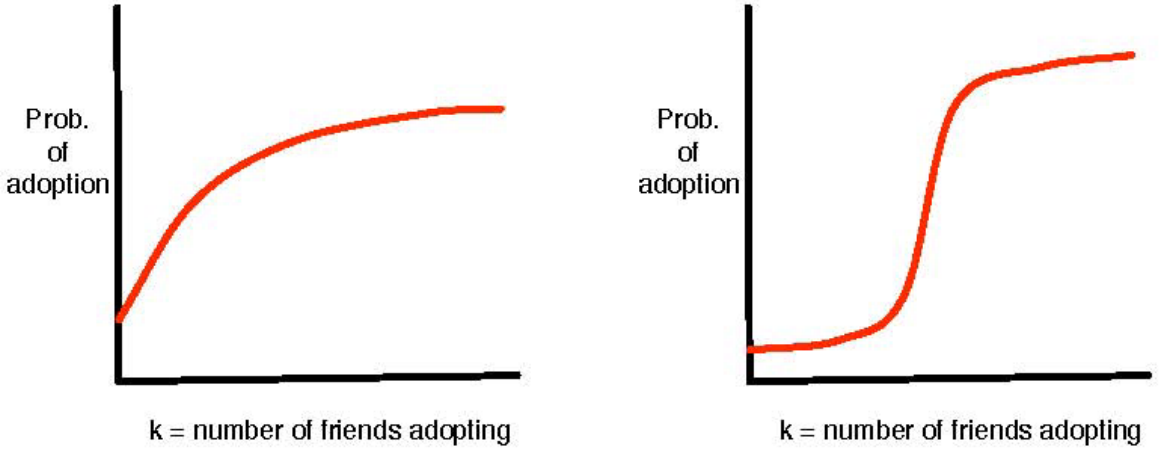
Diffusion, another fundamental social processs:
Behaviors that cascade from node to node like an epidemic.

- News, opinions, rumors, fads, urban legends, ...
- Viral marketing [Domingos-Richardson 2001]
- Public health (e.g. obesity [Christakis-Fowler 2007])
- Cascading failures in financial markets
- Localized collective action: riots, walkouts

Diffusion Curves

Basis for models: Probability of adopting new behavior depends on number of friends who have adopted.

- Bass 1969; Granovetter 1978; Schelling 1978

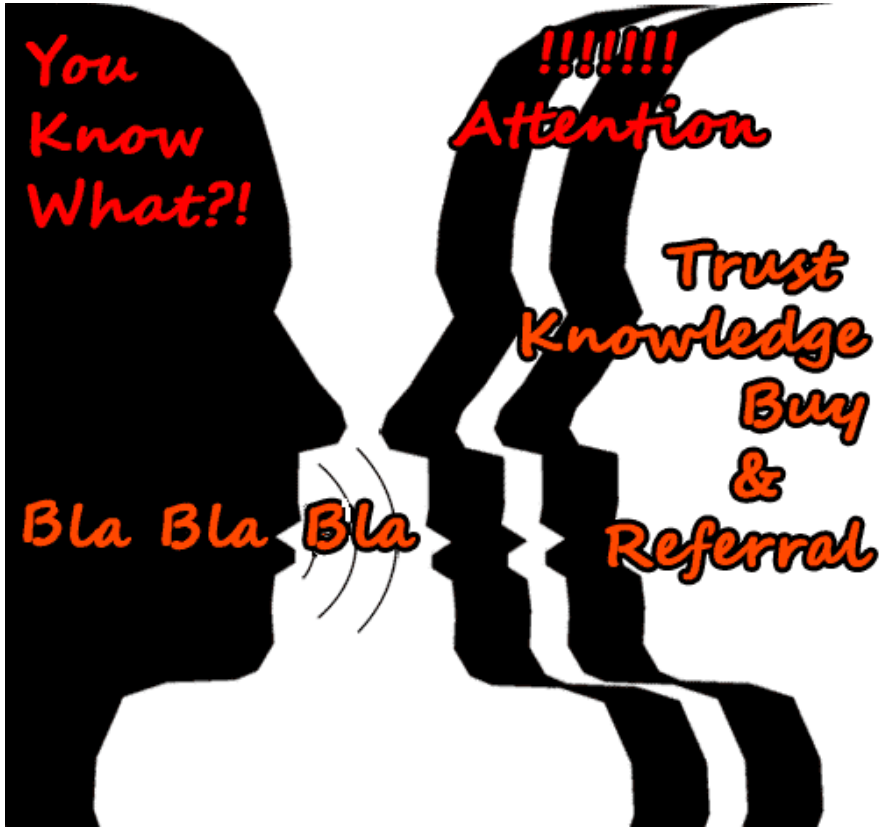


Key issue: qualitative shape of the diffusion curves.

- Diminishing returns? Critical mass?
- Distinction has consequences for models of diffusion at population level.

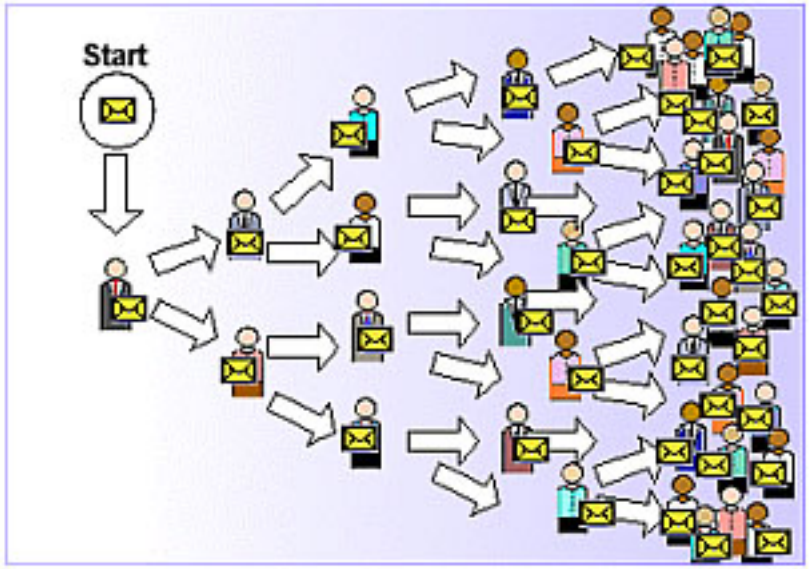
Word of Mouth and Viral Marketing

- We are more influenced by our friends than strangers
- 68% of consumers consult friends and family before purchasing home electronics (Burke 2003)



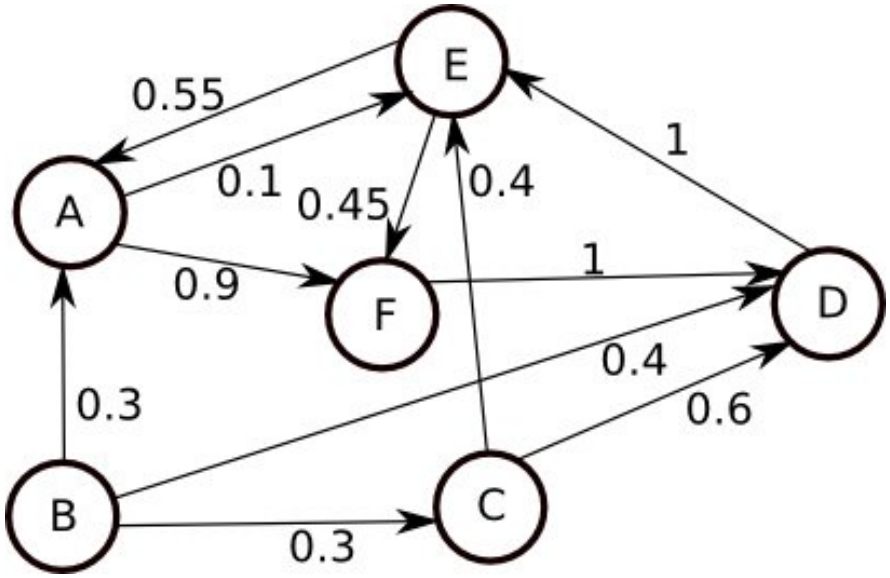
Viral Marketing

- Also known as Target Advertising
- Initiate chain reaction by Word of mouth effect
- Low investments, maximum gain



Viral Marketing as an Optimization Problem

- Given: Network with influence probabilities
- Problem: Select **top-k** users such that by targeting them, the spread of influence is maximized



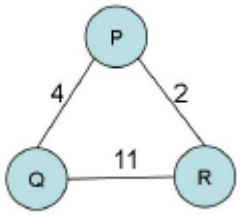
- How to calculate true influence probabilities?

Some Questions

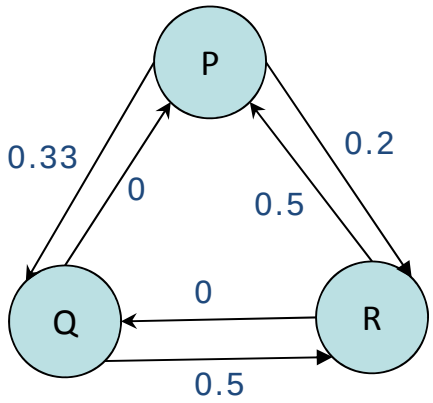
- Where do those influence probabilities come from?
  - Available real world datasets don't have prob.!
- Can we learn those probabilities from available data?
- Previous Viral Marketing studies ignore the effect of time.
  - How can we take time into account?
    - Do probabilities change over time?
  - Can we predict time at which user is most likely to perform an action.
- What users/actions are more prone to influence?

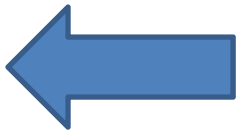
Overview of the proposed framework

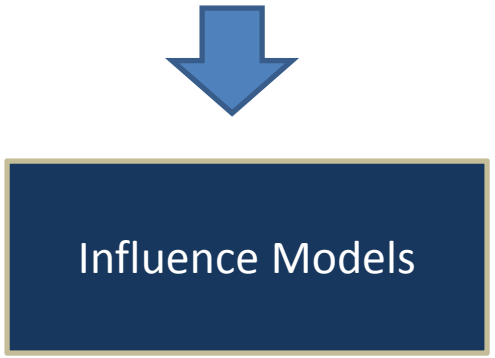
- Input:
  - Social Graph: P and Q become friends at time 4.
  - Action log: User P performs actions a1 at time unit 5.



User	Action	Time
P	a1	5
Q	a1	10
R	a1	15
Q	a2	12
R	a2	14
R	a3	6
P	a3	14







Amit Goyal, Francesco Bonchi, and Laks V.S. Lakshmanan. 2010. Learning influence probabilities in social networks. In *Proceedings of the third ACM international conference on Web search and data mining* (WSDM '10).