Data Mining: Introduction

CS345a: Data Mining
Jure Leskovec
Stanford University

Course Staff

- Instructors:
  - Jure Leskovec
  - Anand Rajaraman
- TAs:
  - Abhishek Gupta
  - Roshan Sumbaly
- Reach us at cs345a-win0910-staff@lists.stanford.edu
- More info on www.stanford.edu/class/cs345a

Requirements

- Homework: 20%
  - Gradiance and other
  - 3 late days for the quarter
  - All homeworks must be handed in
- Project: 40%
  - Start early
  - Takes lots of time
- Final Exam: 40%

Prerequisites

- Basic databases: CS145
- Algorithms:
  - Dynamic programming, basic data structures
- Basic statistics:
  - Moments, typical distributions, regression, …
- Programming:
  - Your choice, but C++/Java will be very useful
- We provide some background, but the class will be fast paced

Class Project

- Software implementation related to course subject matter
- Should involve an original component or experiment
- More later about available data and computing resources

- It's going to be fun and hard work

Possible Projects

- Many past projects have dealt with *collaborative filtering* (advice based on what similar people do)
  - E.g., Netflix Challenge
- Others have dealt with engineering solutions to machine-learning problems
- Lots of interesting project ideas
  - If you can't think of one please come talk to us

Projects: Data & Infrastructure

- Data:
  - Netflix
  - WebBase
  - Wikipedia
  - TREC
  - ShareThis
  - Google
- Infrastructure:
  - Aster Data cluster on Amazon EC2
  - Supports both MapReduce and SQL

Projects: Machine learning

- ML generally requires a large "training set" of correctly classified data:
  - Example: classify Web pages by topic

- Hard to find well-classified data:
  - Open Directory works for page topics, because work is collaborative and shared by many.
  - Other good exceptions?

Projects: Thought

- Many problems require thought:
  1. Tell important pages from unimportant (PageRank)
  2. Tell real news from publicity (how?)
  3. Distinguish positive from negative product reviews (how?)
  4. Feature generation in ML
  5. Etc., etc.

Team Projects

- Working in pairs OK, but …
  1. No more than two per project.
  2. We will expect more from a pair than from an individual.
  3. The effort should be roughly evenly distributed.

Course Outline (1)

- Map-Reduce and Hadoop
- Recommendation systems
  - Collaborative filtering
- Dimensionality reduction
- Finding nearest neighbors
- Finding similar sets
  - Minhashing, Locality-Sensitive hashing
- Clustering
- PageRank and measures of importance in graphs (*link analysis*)
  - Spam detection
  - Topic-specific search

Course Outline (2)

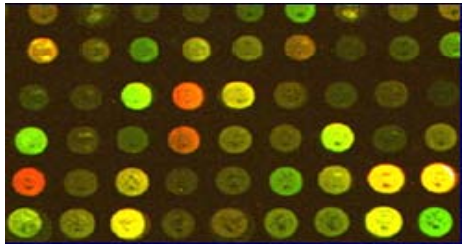
- Large scale machine learning
- Association rules, frequent itemsets
- Extracting structured data (relations) from the Web
- Clustering data
- Graph partitioning
- Spam detection
- Managing Web advertisements
- Mining data streams

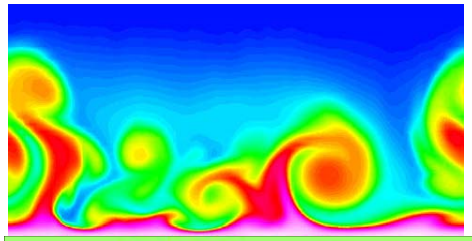
Why Mine Data? Industry

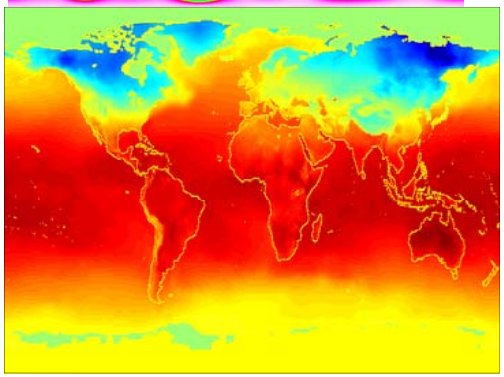
- Lots of data is being collected and warehoused
  - Web data, e-commerce
  - purchases at department/ grocery stores
  - Bank/Credit Card transactions

- Computers are cheap and powerful
- Competitive Pressure is Strong
  - Provide better, customized services for an edge (e.g. in Customer Relationship Management)

Why Mine Data? Science



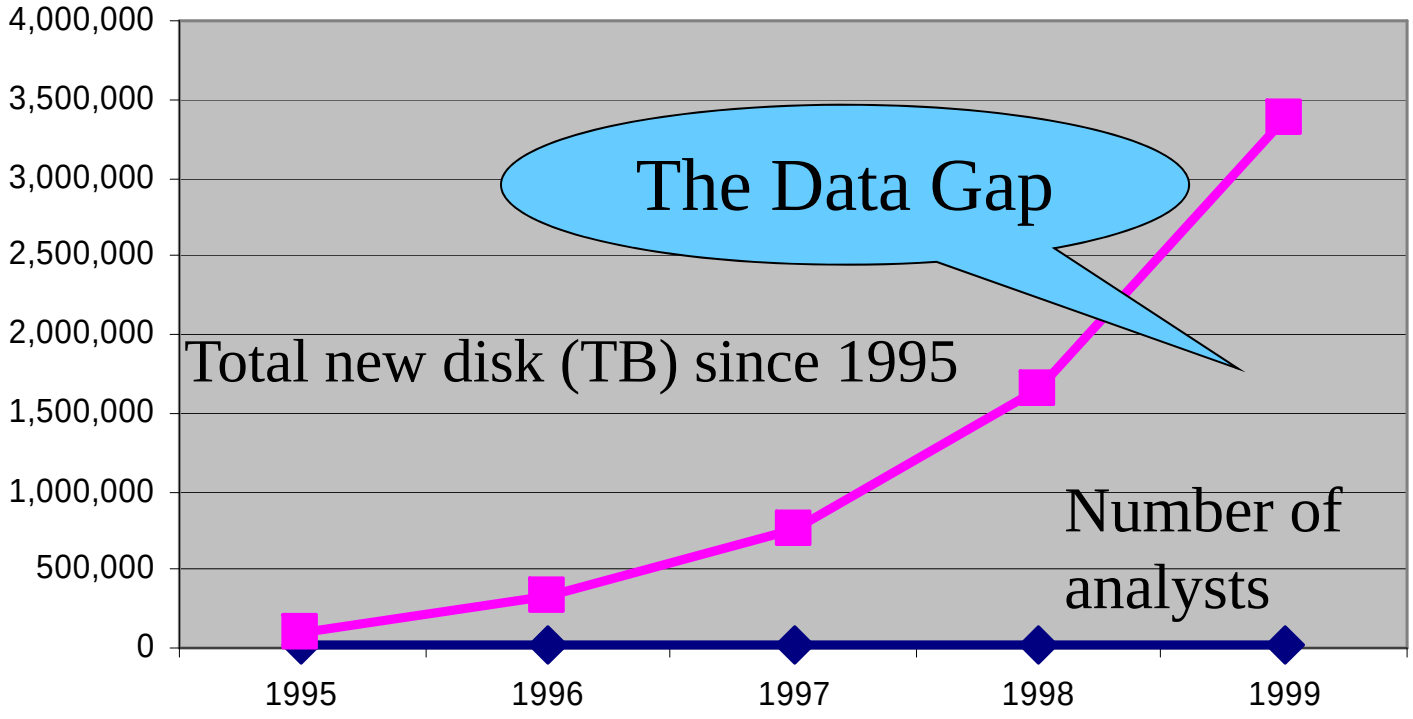




- Data collected and stored at enormous speeds (GB/hour)
  - remote sensors on a satellite
  - telescopes scanning the skies
  - microarrays generating gene expression data
  - scientific simulations generating terabytes of data
- Traditional techniques infeasible for raw data
- Data mining helps scientists
  - in classifying and segmenting data
  - in Hypothesis Formation

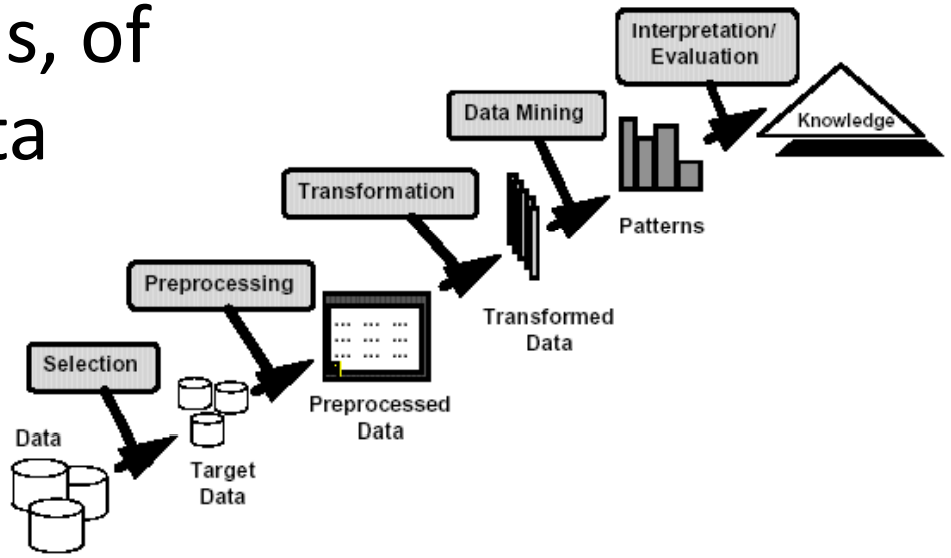
Mining Large Data Sets - Motivation

- There is often information "hidden" in the data that is not readily evident
- Human analysts take weeks to discover useful information
- Much of the data is never analyzed at all



What is Data Mining?

- Many Definitions
  - Non-trivial extraction of implicit, previously unknown and useful information from data
  - Exploration & analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns



Data mining

- Process of semi-automatically analyzing large databases to find patterns that are:
  - valid: hold on new data with some certainty
  - novel: non-obvious to the system
  - useful: should be possible to act on the item
  - understandable: humans should be able to interpret the pattern

Meaningfulness of Answers

- A big data-mining risk is that you will "discover" patterns that are meaningless.

- Bonferroni's principle: (roughly) if you look in more places for interesting patterns than your amount of data will support, you are bound to find crap.

Rhine Paradox – (1)

- A parapsychologist in the 1950's hypothesized that some people had Extra-Sensory Perception
- He devised an experiment where subjects were asked to guess 10 hidden cards – red or blue
- He discovered that almost 1 in 1000 had ESP – they were able to get all 10 right

Rhine Paradox – (2)

- He told these people they had ESP and called them in for another test of the same type
- Alas, he discovered that almost all of them had lost their ESP
- What did he conclude?

- He concluded that you shouldn't tell people they have ESP; it causes them to lose it. ☺

Applications

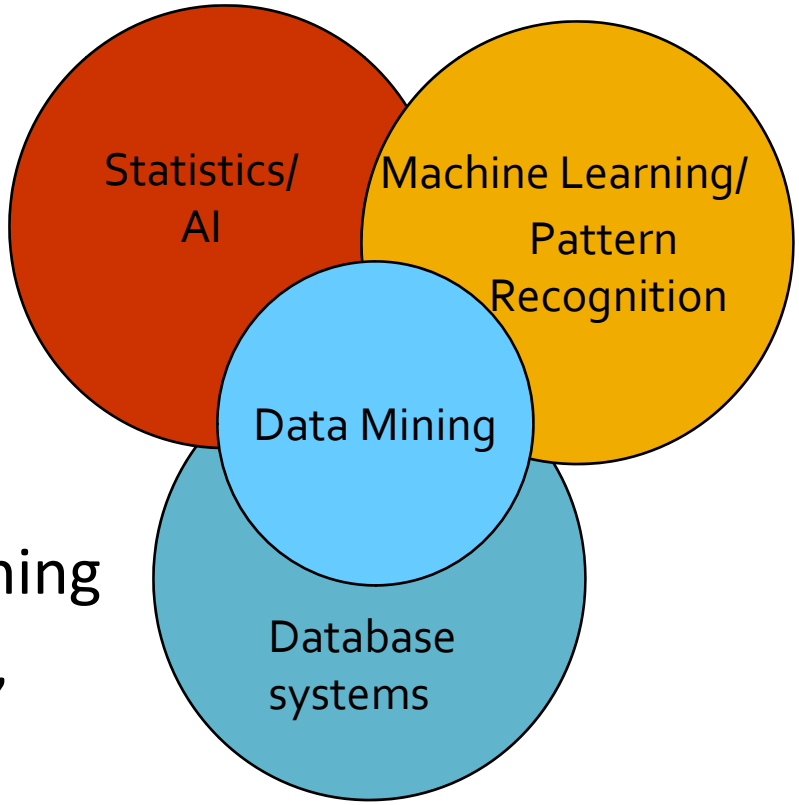
- Banking: loan/credit card approval:
  - predict good customers based on old customers
- Customer relationship management:
  - identify those who are likely to leave for a competitor
- Targeted marketing:
  - identify likely responders to promotions
- Fraud detection: telecommunications, finance
  - from an online stream of event identify fraudulent events
- Manufacturing and production:
  - automatically adjust knobs when process parameter changes

Applications (continued)

- Medicine: disease outcome, effectiveness of treatments
  - analyze patient disease history: find relationship between diseases
- Molecular/Pharmaceutical:
  - identify new drugs
- Scientific data analysis:
  - identify new galaxies by searching for sub clusters
- Web site/store design and promotion:
  - find affinity of visitor to pages and modify layout

Origins of Data Mining

- Overlaps with machine learning, statistics, artificial intelligence, databases, visualization but more stress on
  - scalability of number of features and instances
  - stress on algorithms and architectures whereas foundations of methods and formulations provided by statistics and machine learning
  - automation for handling large, heterogeneous data



Data Mining Tasks

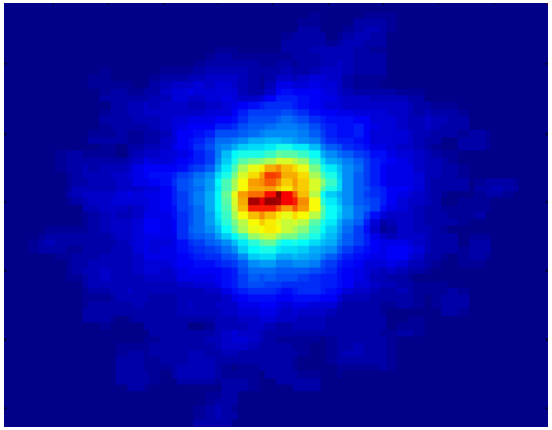
- Prediction Methods
  - Use some variables to predict unknown or future values of other variables.
- Description Methods
  - Find human-interpretable patterns that describe the data.

Data Mining Tasks…

- Classification
- Clustering
- Association Rule Discovery:
- Sequential Pattern Discovery
- Regression
- Anomaly Detection

Classifying Galaxies

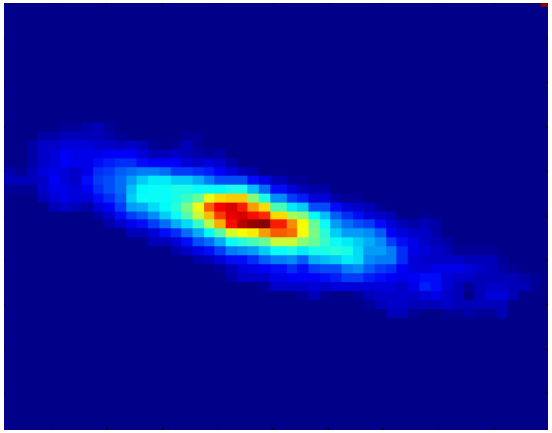
Early



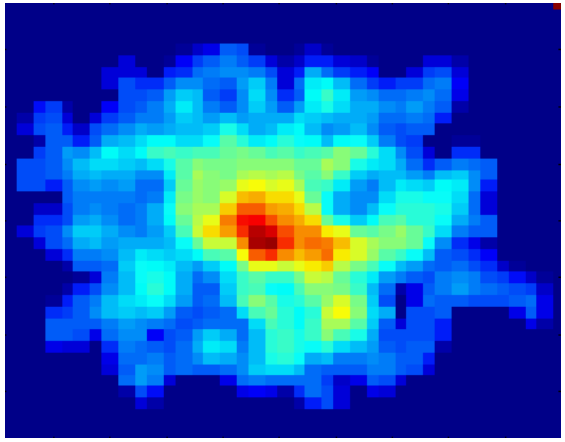
Class:
- Stages of Formation

Attributes:
- Image features,
- Characteristics of light waves received, etc.

Intermediate



Late



Data Size:
- 72 million stars, 20 million galaxies
- Object Catalog: 9 GB
- Image Database: 150 GB

Clustering S&P 500 Stock Data

- Observe Stock Movements
- Cluster them: Stock-{UP/DOWN}
- Similarity Measure:
  - Two points are more similar if the events described by them frequently happen together on the same day.

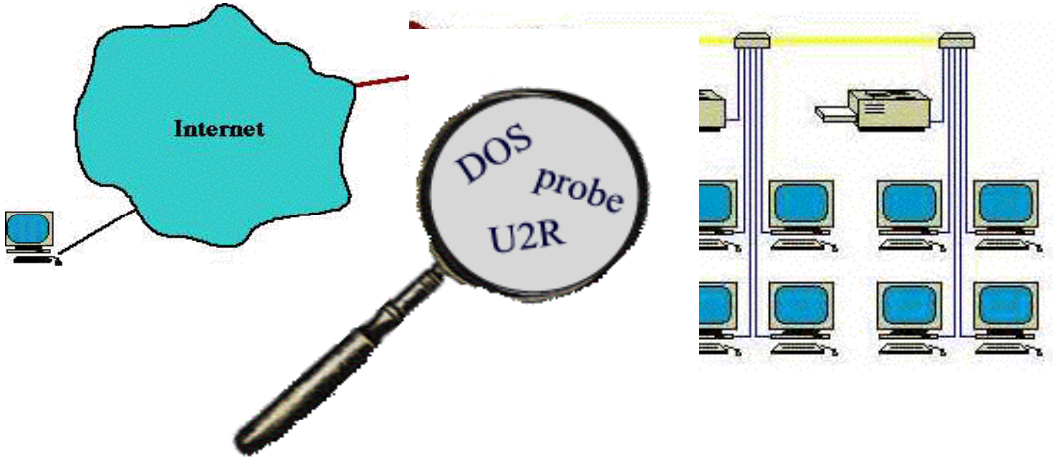
	Discovered Clusters	*Industry Group*
1	Applied-Matl-DOWN,Bay-Network-Down,3-COM-DOWN, Cabletron-Sys-DOWN,CISCO-DOWN,HP-DOWN, DSC-Comm-DOWN,INTEL-DOWN,LSI-Logic-DOWN, Micron-Tech-DOWN,Texas-Inst-Down,Tellabs-Inc-Down, Natl-Semiconduct-DOWN,Oracl-DOWN,SGI-DOWN, Sun-DOWN	**Technology1-DOWN**
2	Apple-Comp-DOWN,Autodesk-DOWN,DEC-DOWN, ADV-Micro-Device-DOWN,Andrew-Corp-DOWN, Computer-Assoc-DOWN,Circuit-City-DOWN, Compaq-DOWN, EMC-Corp-DOWN, Gen-Inst-DOWN, Motorola-DOWN,Microsoft-DOWN,Scientific-Atl-DOWN	Technology2-DOWN
3	Fannie-Mae-DOWN,Fed-Home-Loan-DOWN, MBNA-Corp-DOWN,Morgan-Stanley-DOWN	Financial-DOWN
4	Baker-Hughes-UP,Dresser-Inds-UP,Halliburton-HLD-UP, Louisiana-Land-UP,Phillips-Petro-UP,Unocal-UP, Schlumberger-UP	Oil-UP

Collaborative Filtering

- Given database of user preferences, predict preference of new user
- Example:
  - Predict what new movies you will like based on
    - your past preferences
    - others with similar past preferences
    - their preferences for the new movies
- Example:
  - Predict what books/CDs a person may want to buy
  - (and suggest it, or give discounts to tempt customer)

Anomaly Detection

- Detect significant deviations from normal behavior
- Applications:
  - Credit Card Fraud Detection
  - Network Intrusion Detection



Association Rule Discovery

- Supermarket shelf management.
  - Goal: To identify items that are bought together by sufficiently many customers.
  - Approach: Process the point-of-sale data collected with barcode scanners to find dependencies among items.
  - A classic rule --
    - If a customer buys diaper and milk, then he is likely to buy beer.
    - So, don't be surprised if you find six-packs stacked next to diapers!

TID	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Rules Discovered:

{Milk} --> {Coke}

{Diaper, Milk} --> {Beer}

Some Success Stories

- Network intrusion detection using a combination of sequential rule discovery and classification tree on 4 GB DARPA data
  - Won over (manual) knowledge engineering approach
  - http://www.cs.columbia.edu/~sal/JAM/PROJECT/ provides good detailed description of the entire process
- Major US bank: Customer attrition prediction
  - Segment customers based on financial behavior: 3 segments
  - Build attrition models for each of the 3 segments
  - 40-50% of attritions were predicted == factor of 18 increase
- Targeted credit marketing: major US banks
  - find customer segments based on 13 months credit balances
  - build another response model based on surveys
  - increased response 4 times – 2%

Challenges of Data Mining

- Scalability
- Dimensionality
- Complex and Heterogeneous Data
- Data Quality
- Data Ownership and Distribution
- Privacy Preservation
- Streaming Data

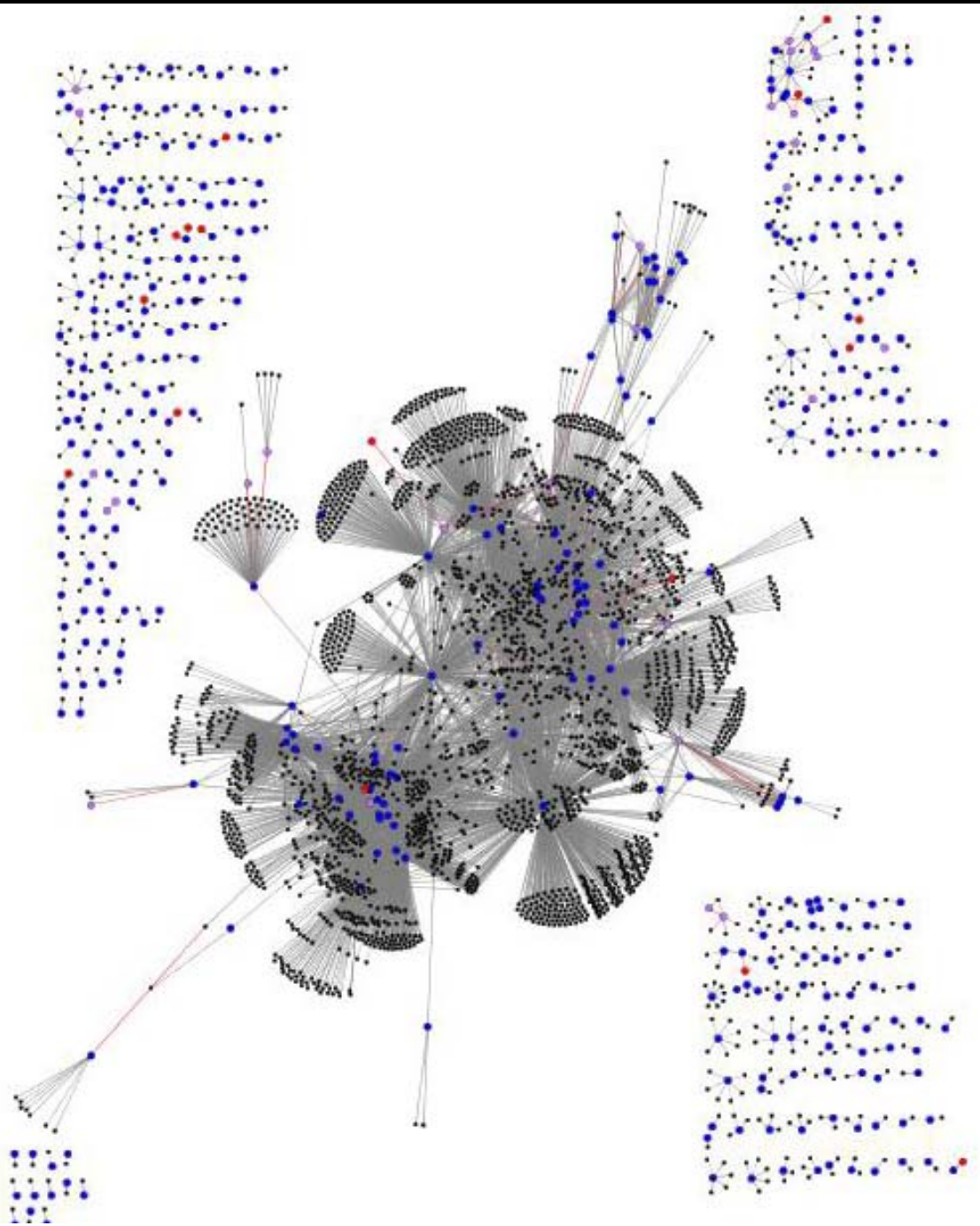
[Leskovec et al., TWEB '07]

Example: Viral Marketing

- Senders and followers of recommendations receive discounts on products

- Recommendations are made to any number of people at the time of purchase
- Only the recipient who buys first gets a discount

Product recommendation network



- purchase following a recommendation
- customer recommending a product
- customer not buying a recommended product

Viral marketing data

- Large online retailer (June 2001 to May 2003)
- 15,646,121 recommendations
- 3,943,084 distinct customers
- 548,523 products recommended
- 99% of them belonging 4 main product groups:
  - books
  - DVDs
  - music
  - VHS

Data attributes

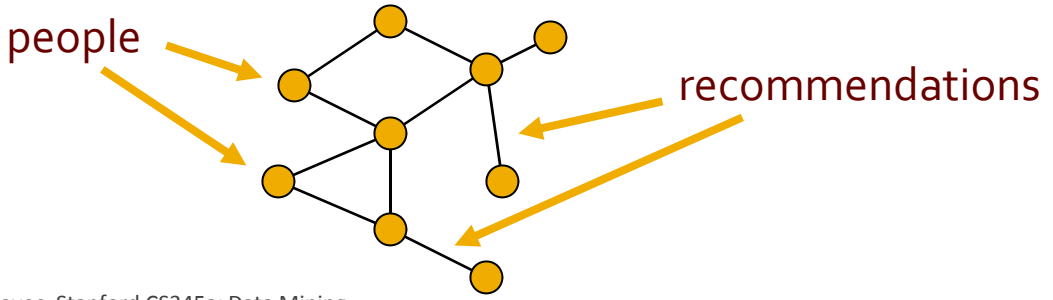
- Recommendations
  - sender (shadowed)
  - recipient (shadowed)
  - recommendation time
  - buy bit
  - purchase time
  - product price
- Additional product info (from the retailer's website)
  - categories
  - reviews
  - ratings

Product group features

- What role does the product category play?

	products	customers	recommenda-tions	edges	buy + get discount	buy + no discount
Book	103,161	**2,863,977**	5,741,611	2,097,809	**65,344**	**17,769**
DVD	**19,829**	805,285	**8,180,393**	962,341	**17,232**	**58,189**
Music	**393,598**	794,148	1,443,847	585,738	7,837	2,739
Video	26,131	239,583	280,270	160,683	909	467
Full	542,719	3,943,084	15,646,121	3,153,676	91,322	79,164

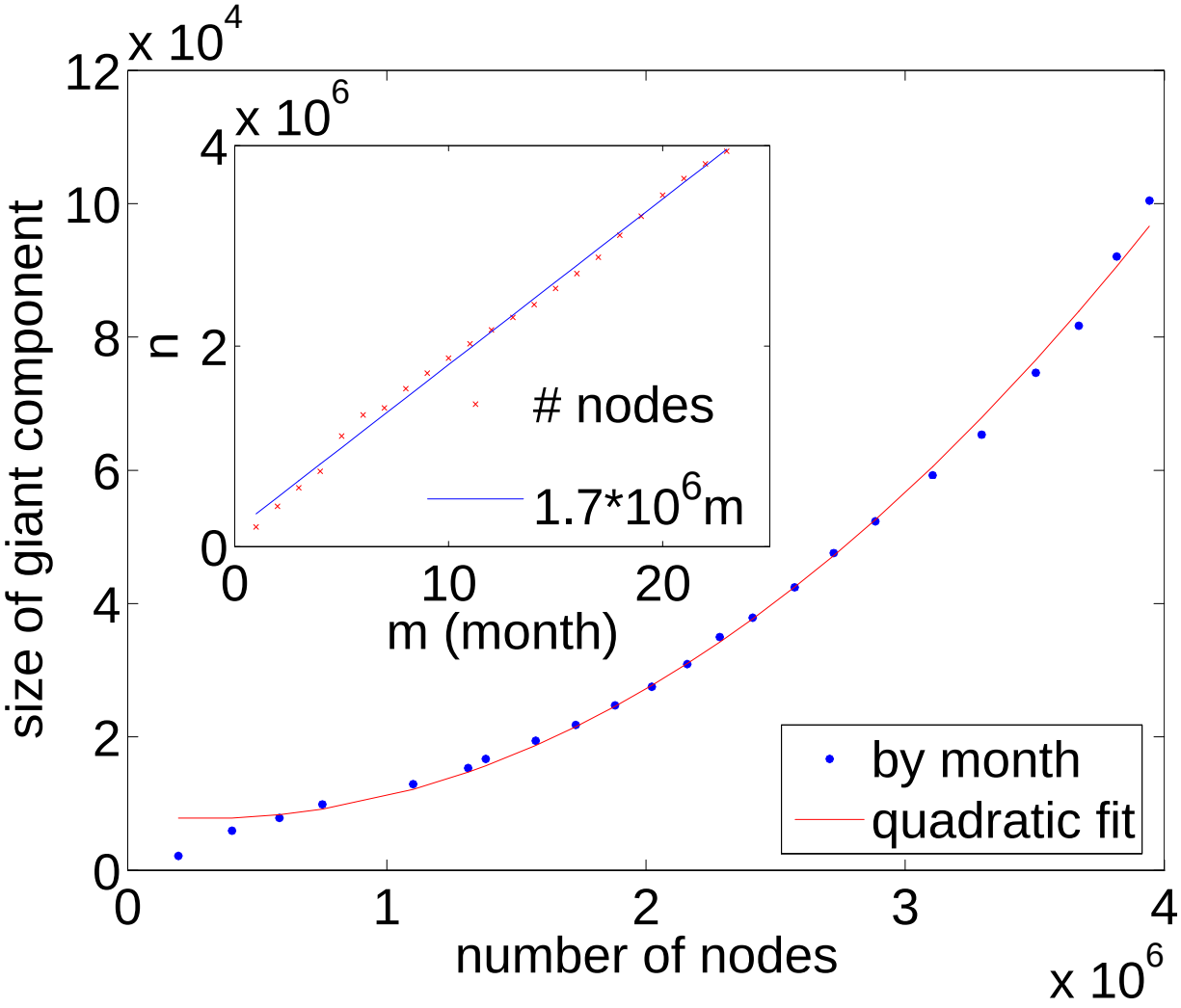
high
low



Observations on product groups

- There are relatively few DVD titles, but DVDs account for ~ 50% of recommendations.
- recommendations per person
  - DVD: 10
  - books and music: 2
  - VHS: 1
- recommendations per purchase
  - books: 69
  - DVDs: 108
  - music: 136
  - VHS: 203
- Overall there are 3.69 recommendations per node on 3.85 different products.
- Music recommendations reached about the same number of people as DVDs but used only 1/5 as many recommendations
- Book recommendations reached by far the most people – 2.8 million.
- All networks have a very small number of unique edges. For books, videos and music the number of unique edges is smaller than the number of nodes – the networks are highly disconnected

Adoption of viral marketing program

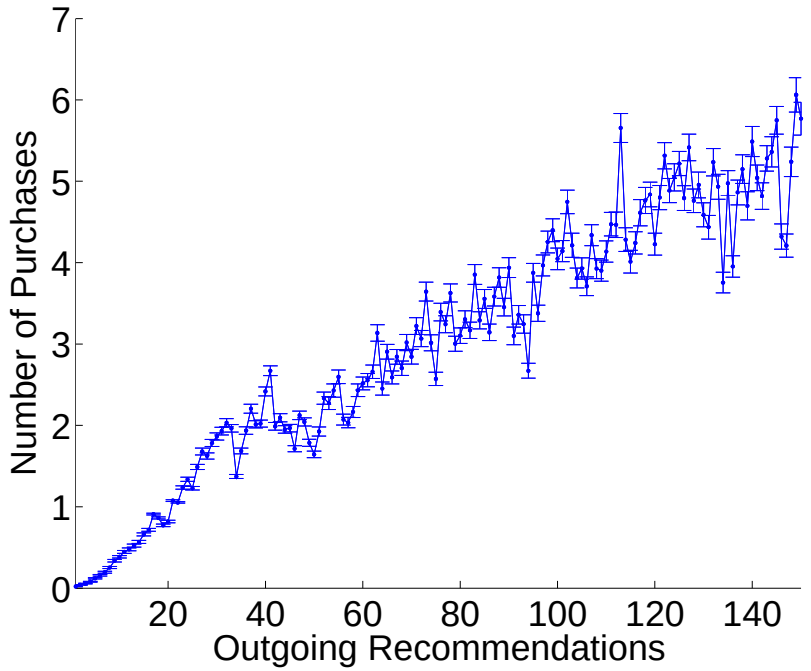


Viral marketing program is not spreading virally

- 94% of users make first recommendation without having received one previously
- linear growth: ~ 165,000 new users added each month
- size of giant connected component increases from 1% to 2.5% of the network (100,420 users) – small!
- some sub-communities are better connected
  - 24% out of 18,000 users for westerns on DVD
  - 26% of 25,000 for classics on DVD
  - 19% of 47,000 for anime (Japanese animated film) on DVD
- others are just as disconnected
  - 3% of 180,000 home and gardening
  - 2-7% for children's and fitness DVDs

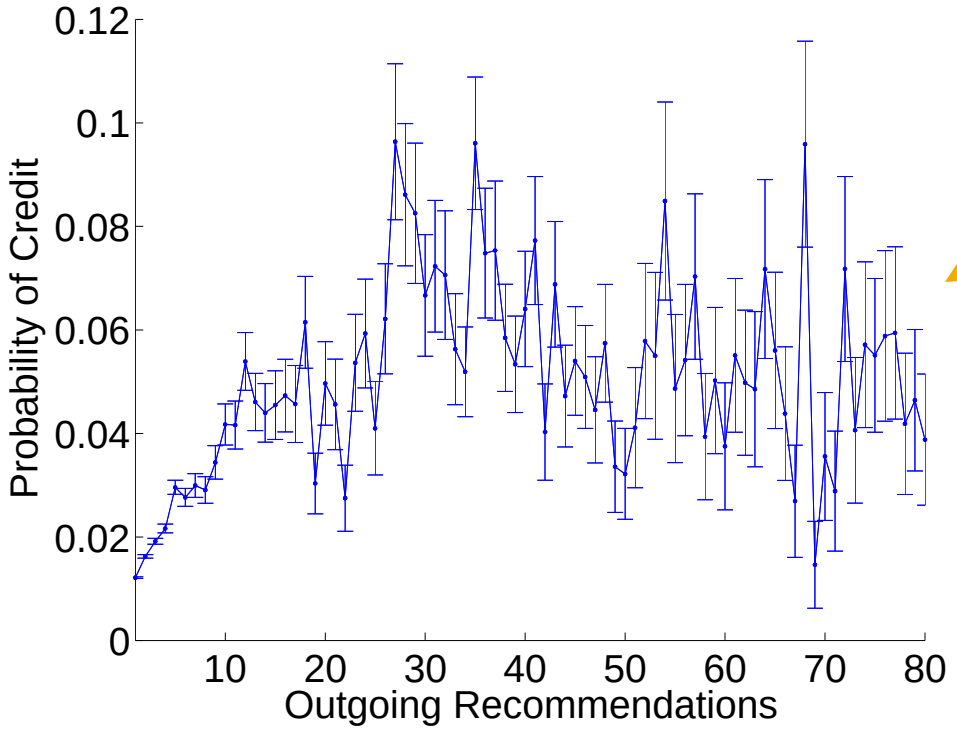
Network effects

- Does sending more recommendations influence more purchases?



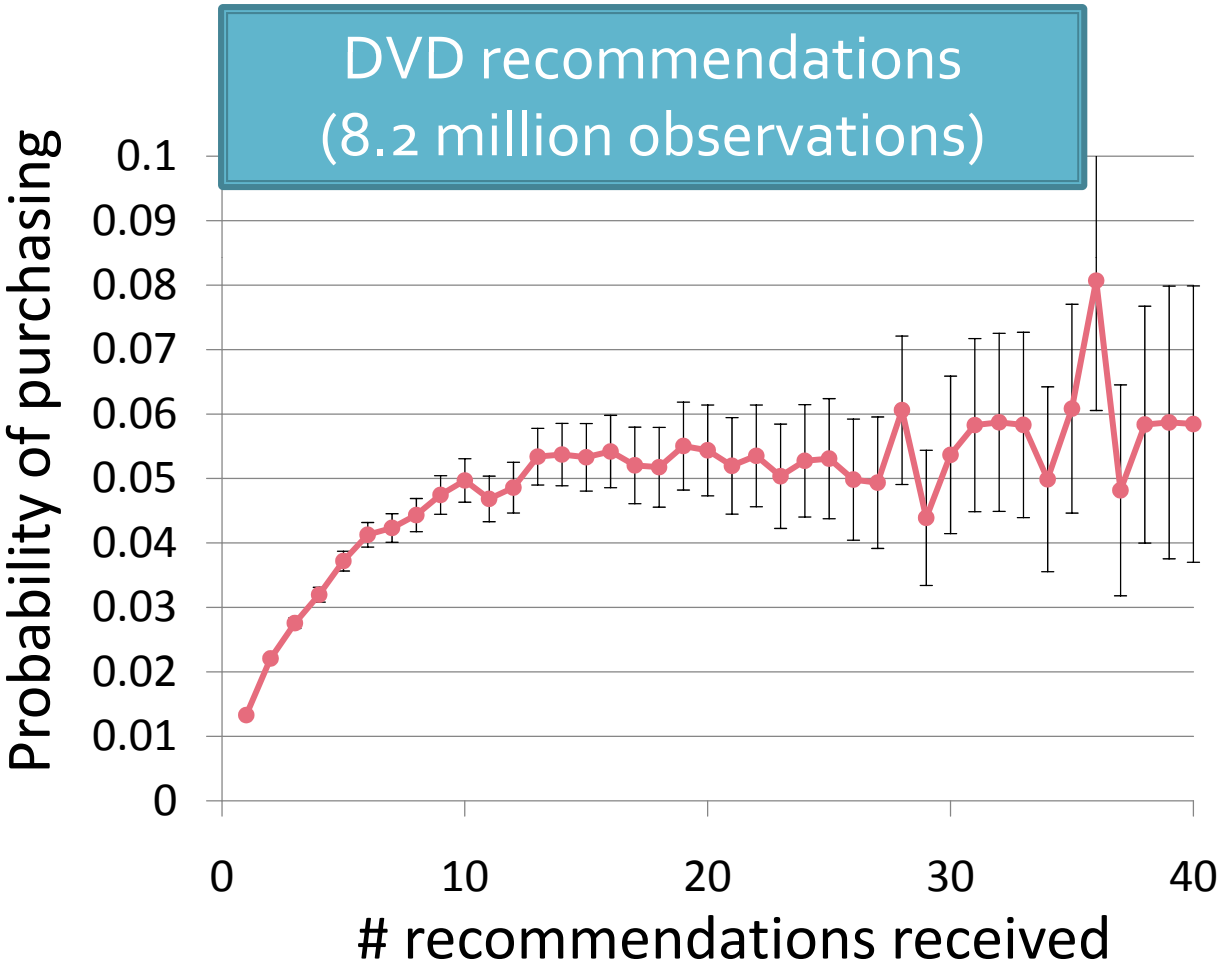
The probability that the sender gets a credit with increasing numbers of recommendations

- consider whether sender has at least one successful recommendation
- controls for sender getting credit for purchase that resulted from others recommending the same product to the same person



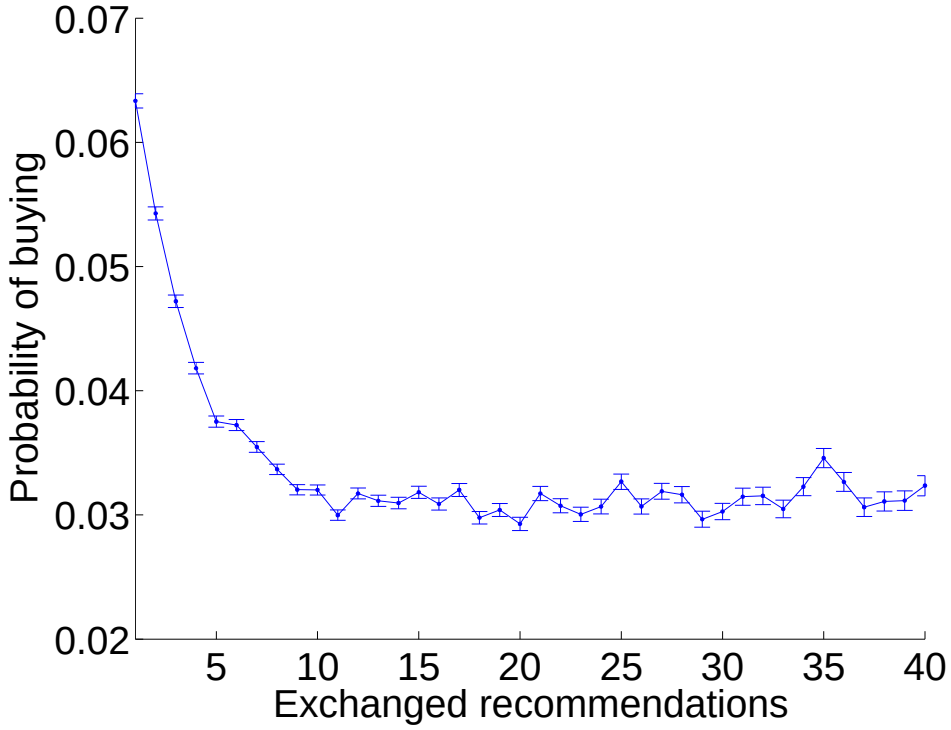
probability of receiving a credit levels off for DVDs

Recommendation response



Recommendation effectiveness

- Effectiveness of subsequent recommendations?
  - Multiple recommendations between two individuals weaken the impact of the bond on purchases



Success by book category

- Consider successful recommendations in terms of
  - av. # senders of recommendations per book category
  - av. # of recommendations accepted
- books overall have a 3% success rate
  - (2% with discount, 1% without)
- Lower than average success rate
  - fiction
    - romance (1.78), horror (1.81)
    - teen (1.94), children's books (2.06)
    - comics (2.30), sci-fi (2.34), mystery and thrillers (2.40)
  - nonfiction
    - sports (2.26)
    - home & garden (2.26)
    - travel (2.39)
- Higher than average success rate
  - professional & technical
    - medicine (5.68)
    - professional & technical (4.54)
    - engineering (4.10), science (3.90), computers & internet (3.61)
    - law (3.66), business & investing (3.62)

Professional & organized contexts

- Professional & technical book recommendations are more often accepted
- Some organized contexts other than professional also have higher success rate, e.g. religion
  - overall success rate 3.13%
  - Christian themed books
    - Christian living and theology (4.7%)
    - Bibles (4.8%)
  - not-as-organized religion
    - new age (2.5%)
    - occult spirituality (2.2%)
- Well organized hobbies
  - books on orchids recommended successfully twice as often as books on tomato growing

Predicting recommendation success

Variable	transformation	Coefficient
const		-0.940 ***
# recommendations	ln(r)	0.426 ***
# senders	$\ln(n_s)$	-0.782 ***
# recipients	$\ln(n_r)$	-1.307 ***
product price	ln(p)	0.128 ***
# reviews	ln(v)	-0.011 ***
avg. rating	ln(t)	-0.027 *
R^2		**0.74**

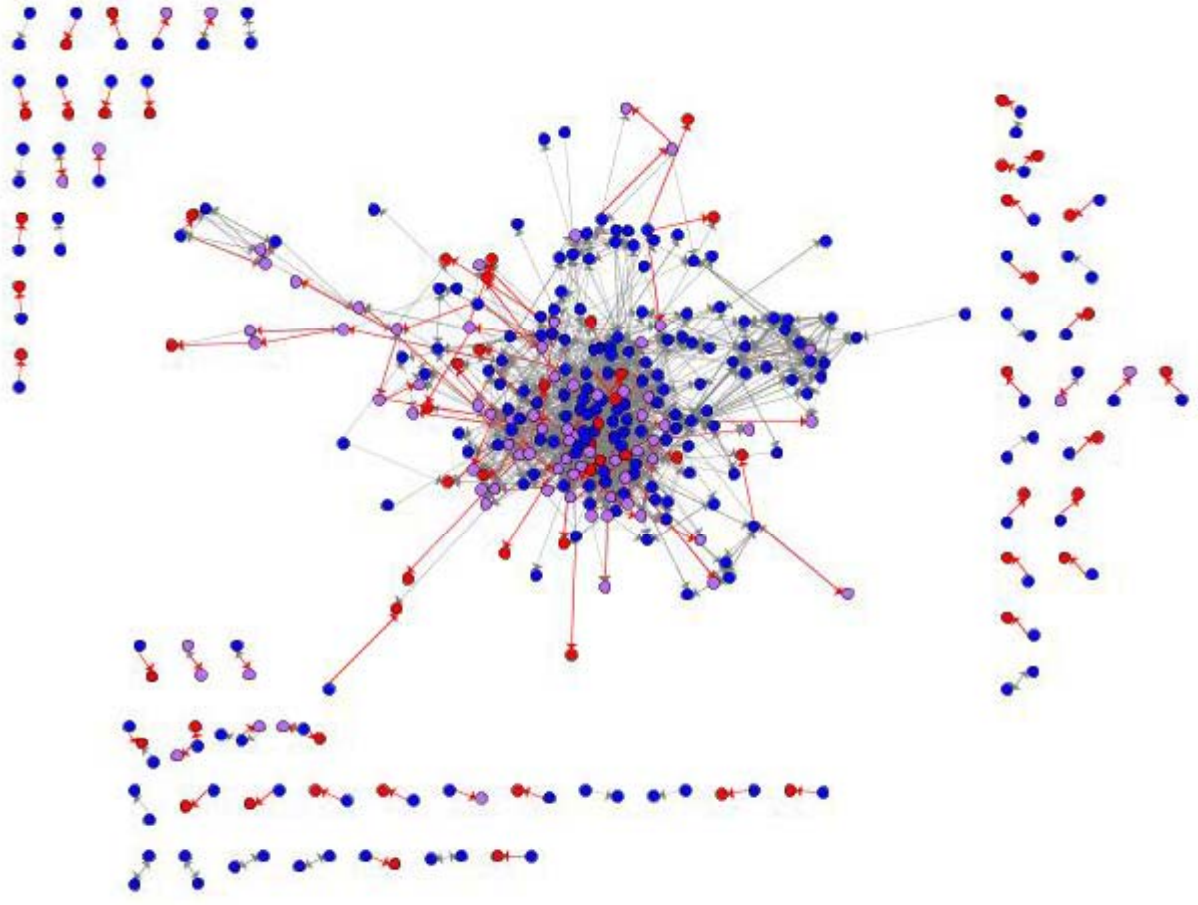
significance at the 0.01 (***), 0.05 (**) and 0.1 (*) levels

Anime

- 47,000 customers responsible for the 2.5 out of 16 million recommendations in the system
- 29% success rate per recommender of an anime DVD
- Giant component covers 19% of the nodes
- Overall, recommendations for DVDs are more likely to result in a purchase (7%), but the anime community stands out

DVD recommendations

- Three colors: blue, white & red
- showing purchasers only



Products suited for Viral Marketing

- Small community
  - few reviews, senders, and recipients
  - but sending more recommendations helps
- Pricey products
- Rating doesn't play as much of a role

Viral Marketing: Consequences

Observations for diffusion models:

- purchase decision more complex than threshold or simple infection
- influence saturates as the number of contacts expands
- links user effectiveness if they are overused

Conditions for successful recommendations:

- professional and organizational contexts
- discounts on expensive items
- small, tightly knit communities