Graph mining (CS 15-826)

Jure Leskovec
http://www.cs.cmu.edu/jure

CMU 15-826 Jure Leskovec (jure@cs.cmu.edu) 1

Outline

- Patterns in real-world graphs
- Modeling the evolution and generation of real graphs
- Graph mining at work:
  - Influence propagation
  - Patterns (bonus material)

CMU 15-826 Jure Leskovec (jure@cs.cmu.edu) 2

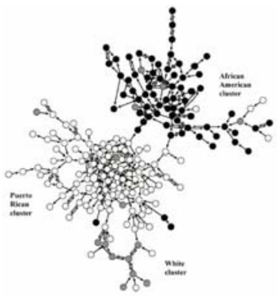
Modeling Graphs over Time

Joint work with:
Jon Kleinberg, Cornell
Deepay Chakrabarti, Yahoo
Christos Faloutsos, CMU

CMU 15-826 Jure Leskovec (jure@cs.cmu.edu) 3

Introduction

- What can we do with graphs?
  - What patterns or “laws” hold for most real-world graphs?
  - How do the graphs evolve over time?
  - Can we generate synthetic but “realistic” graphs?



“Needle exchange” networks of drug users

Evolution of the Graphs

- How do graphs evolve over time?
- Conventional Wisdom:
  - Constant average degree: the number of edges grows linearly with the number of nodes
  - Slowly growing diameter: as the network grows the distances between nodes grow
- Our findings:
  - Densification Power Law: networks are becoming denser over time
  - Shrinking Diameter: diameter is decreasing as the network grows

Outline

- General patterns and generators
- Graph evolution – Observations
  - Densification Power Law
  - Shrinking Diameters
- Proposed explanation
  - Community Guided Attachment
  - Forest Fire Model
- Proposed graph generation model
  - Kronecker Graphs
- Conclusion and Open questions

Outline

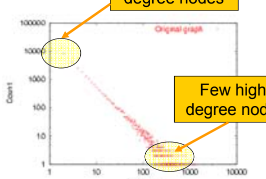
- General patterns and generators
- Graph evolution – Observations
  - Densification Power Law
  - Shrinking Diameters
- Proposed explanation
  - Community Guided Attachment
  - Forest Fire Model
- Proposed graph generation model
  - Kronecker Graphs
- Conclusion

CMU 15-826 Jure Leskovec (jure@cs.cmu.edu) 7

Static Graph Patterns (1)

- Power Law

Internet in December 1998



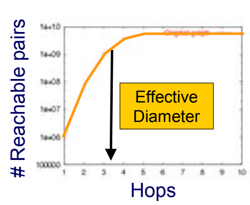
log(Count) vs. log(Degree)

$Y=a*X^b$

CMU 15-826 Jure Leskovec (jure@cs.cmu.edu) 8

Static Graph Patterns (2)

- Small-world [Watts, Strogatz]++
  - 6 degrees of separation
  - Small diameter
- Effective diameter:
  - Distance at which 90% of pairs of nodes are reachable



Epinions who-trusts-whom social network

CMU 15-826 Jure Leskovec (jure@cs.cmu.edu) 9

Patterns Hold in Many Graphs

- All these patterns can be observed in many real life graphs:
  - World wide web [Barabasi]
  - On-line communities [Holme, Edling, Liljeros]
  - Who call whom telephone networks [Cortes]
  - Autonomous systems [Faloutsos, Faloutsos, Faloutsos]
  - Internet backbone – routers [Faloutsos, Faloutsos, Faloutsos]
  - Movie – actors [Barabasi]
  - Science citations [Leskovec, Kleinberg, Faloutsos]
  - Co-authorship [Leskovec, Kleinberg, Faloutsos]
  - Sexual relationships [Liljeros]
  - Click-streams [Chakrabarti]

Graph models: Random Graphs

- Question: How can we generate a realistic graph?
  - given the number of nodes N and edges E
- Random graph [Erdos & Renyi, 60s]:
  - Pick 2 nodes at random and link them
  - Nice and simple model
  - Does not obey Power laws
  - No community structure

Graph models: Preferential attachment

- Preferential attachment [Albert & Barabasi, 99]:
  - Add a new node, create M out-links
  - Probability of linking a node is proportional to its degree
- Examples:
  - Citations: new citations of a paper are proportional to the number it already has
- “Rich get richer” phenomena
- Explains power-law degree distributions
- But, all nodes have equal (constant) out-degree

Graph models: Copying model

- Copying model [Kleinberg, Kumar, Raghavan, Rajagopalan and Tomkins, 99]:
  - Add a node and choose the number of edges to add
  - Choose a random vertex and "copy" its links (neighbors)
- Generates power-law degree distributions
- Generates communities

Why is all this important?

- Gives insight into the graph formation process:
  - Anomaly detection – abnormal behavior, evolution
  - Predictions – predicting future from the past
  - Simulations of new algorithms where real graphs are hard/impossible to collect
  - Graph sampling – many real world graphs are too large to deal with
  - "What if" scenarios

Outline

- General patterns and generators
- Graph evolution – Observations
  - Densification Power Law
  - Shrinking Diameters
- Proposed explanation
  - Community Guided Attachment
  - Forest Fire Model
- Proposed graph generation model
  - Kronecker Graphs
- Conclusion

Temporal Evolution of the Graphs

- $N(t)$ … nodes at time t
- $E(t)$ … edges at time t
- Suppose that
  $N(t+1) = 2 * N(t)$
- Q: what is your guess for
  $E(t+1) = \cancel{2} * E(t)$
- A: over-doubled!
  - But obeying the Densification Power Law

CMU 15-826 Jure Leskovec (jure@cs.cmu.edu) 16

Temporal Evolution of the Graphs

- Densification Power Law
  - networks are becoming denser over time
  - the number of edges grows faster than the number of nodes – average degree is increasing

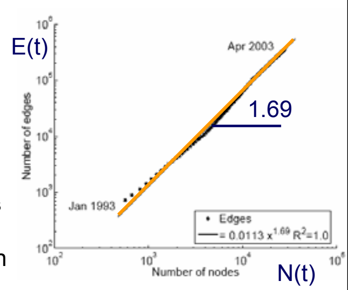
$$\boxed{E(t) \propto N(t)^a} \quad \text{or equivalently} \quad \frac{\log(E(t))}{\log(N(t))} = const$$

a … densification exponent: 1 ≤ a ≤ 2:
  - a=1: linear growth – constant out-degree (assumed in the literature so far)
  - a=2: quadratic growth – clique

CMU 15-826 Jure Leskovec (jure@cs.cmu.edu) 17

Densification – Physics Citations

- Citations among physics papers
- 1992:
  - 1,293 papers, 2,717 citations
- 2003:
  - 29,555 papers, 352,807 citations
- For each month M, create a graph of all citations up to month M



CMU 15-826 Jure Leskovec (jure@cs.cmu.edu) 18

Graph Densification – Summary

- The traditional constant out-degree assumption does not hold
- Instead:

$$E(t) \propto N(t)^a$$

- the number of edges grows faster than the number of nodes – average degree is increasing

Outline

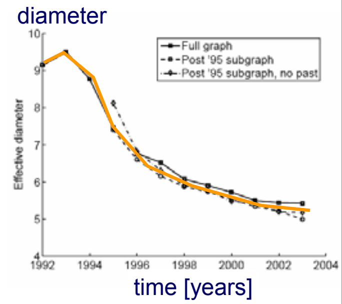
- General patterns and generators
- Graph evolution – Observations
  - Densification Power Law
  - Shrinking Diameters
- Proposed explanation
  - Community Guided Attachment
  - Forest Fire Model
- Proposed graph generation model
  - Kronecker Graphs
- Conclusion

Evolution of the Diameter

- Prior work on Power Law graphs hints at Slowly growing diameter:
  - diameter ~ O(log N)
  - diameter ~ O(log log N)
- What is happening in real data?
- Diameter shrinks over time
  - As the network grows the distances between nodes slowly decrease

Diameter – ArXiv citation graph

- Citations among physics papers
- 1992 –2003
- One graph per year



CMU 15-826 Jure Leskovec (jure@cs.cmu.edu) 22

Patterns Hold in Many Graphs

- Densification and Shrinking diameter can be observed in many real life graphs:
  - Science citations
  - Patent citations
  - Autonomous systems
  - Movie – actors
  - Co-authorship
  - Authors – papers
  - Recommendation networks
  - Email networks

CMU 15-826 Jure Leskovec (jure@cs.cmu.edu) 23

Outline

- General patterns and generators
- Graph evolution – Observations
  - Densification Power Law
  - Shrinking Diameters
- Proposed explanation
  - Community Guided Attachment
  - Forest Fire Model
- Proposed graph generation model
  - Kronecker Graphs
- Conclusion

CMU 15-826 Jure Leskovec (jure@cs.cmu.edu) 24

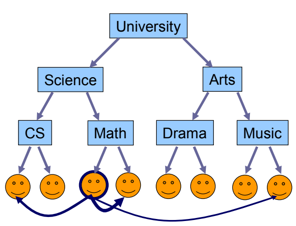
Densification – Possible Explanation

- Existing graph generation models do not capture the Densification Power Law and Shrinking diameters
- Can we find a simple model of local behavior, which naturally leads to observed phenomena?
- Yes! We present 2 models:
  - Community Guided Attachment – obeys Densification
  - Forest Fire model – obeys Densification, Shrinking diameter (and Power Law degree distribution)

CMU 15-826 Jure Leskovec (jure@cs.cmu.edu) 25

Community structure

- Let's assume the community structure
- One expects many within-group friendships and fewer cross-group ones



- How hard is it to cross communities?

Self-similar university community structure

CMU 15-826 Jure Leskovec (jure@cs.cmu.edu) 26

Main Assumption

- Assume the cross-community linking probability of nodes at tree-distance h is scale-free
- Then the cross-community linking probability is:

$$\boxed{f(h) = c^{-h}}$$

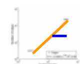
where: $c \geq 1$ … the *Difficulty constant*
h … tree-distance

CMU 15-826 Jure Leskovec (jure@cs.cmu.edu) 27

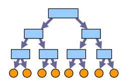
Densification Power Law (2)

- Theorem: The Community Guided Attachment leads to Densification Power Law with exponent

$$\boxed{a = 2 - \log_b(c)}$$



- a … densification exponent $E(t) \propto N(t)^a$
- b … community tree branching factor
- c … difficulty constant, $1 \le c \le b$



Difficulty Constant

- Theorem: $\boxed{a = 2 - \log_b(c)}$
- Gives any non-integer Densification exponent
- If c = 1: easy to cross communities
  - Then: a=2, quadratic growth of edges – near clique
- If c = b: hard to cross communities
  - Then: a=1, linear growth of edges – constant out-degree

Outline

- General patterns and generators
- Graph evolution – Observations
  - Densification Power Law
  - Shrinking Diameters
- Proposed explanation
  - Community Guided Attachment
  - Forest Fire Model
- Proposed graph generation model
  - Kronecker Graphs
- Conclusion

“Forest Fire” model – Wish List

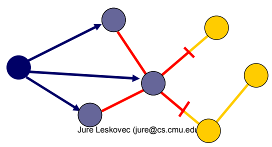
- We want:
  - no explicit Community structure
  - shrinking diameters
- And a bit of:
  - “Rich get richer” attachment process, to get heavy-tailed in-degrees
  - “Copying” model, to lead to communities
  - Community Guided Attachment, to produce Densification Power Law

“Forest Fire” model – Intuition

- How do authors identify references?
  1. Find first paper and cite it
  2. Follow a few citations, make citations
  3. Continue recursively
  4. From time to time use bibliographic tools (e.g. CiteSeer) and chase back-links

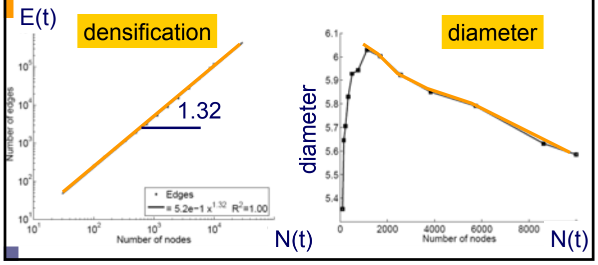
“Forest Fire” – Example

- A node arrives
- Randomly chooses an “ambassador”
- Starts burning nodes (with probability *p*) and adds links to burned nodes
- “Fire” spreads recursively



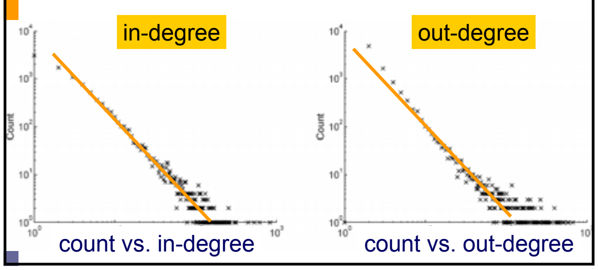
Forest Fire in Action (1)

- Simulation experiments
- Forest Fire generates graphs that Densify and have Shrinking Diameter



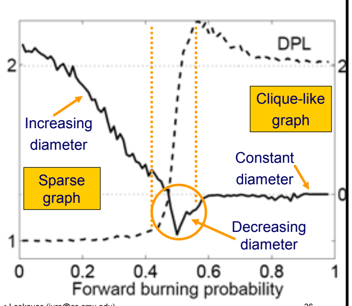
Forest Fire in Action (2)

- Forest Fire also generates graphs with heavy-tailed degree distribution



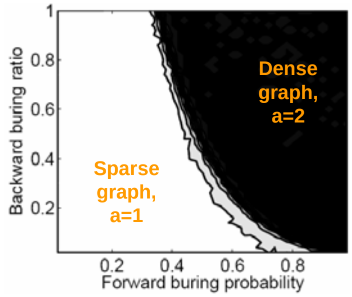
Forest Fire: Parameter Space

- Fix backward probability p_b and vary forward burning probability p
- We observe a sharp transition between sparse and clique-like graphs
- Sweet spot is very narrow



Forest Fire: Full Parameter Space

- We observe a sharp transition between sparse and clique-like graphs
- Sweet spot is very narrow



CMU 15-826 Jure Leskovec (jure@cs.cmu.edu) 37

Recap

- We have seen static graph patterns
- We observed two new temporal graph patterns
  - Densification Power Law
  - Shrinking Diameter
- We found intuitive explanation
- Question: How can we generate a realistic graph?
  - given the number of nodes N and edges E

CMU 15-826 Jure Leskovec (jure@cs.cmu.edu) 38

Outline

- General patterns and generators
- Graph evolution – Observations
  - Densification Power Law
  - Shrinking Diameters
- Proposed explanation
  - Community Guided Attachment
  - Forest Fire Model
- Proposed graph generation model
  - Kronecker Graphs
- Conclusion

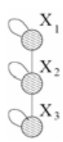
CMU 15-826 Jure Leskovec (jure@cs.cmu.edu) 39

Problem Definition

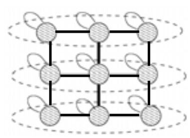
- Given a growing graph with nodes N_1, N_2, …
- Generate a realistic sequence of graphs that will obey all the patterns:
  - Static Patterns
    - Power Law Degree Distribution
    - Small Diameter
    - Power Law Eigenvalue and Eigenvector Distribution
  - Temporal Patterns
    - Densification Power Law
    - Shrinking/Constant Diameters
  - And ideally we would like to prove them

Recursive Graph Generation

- There are many obvious (but wrong) ways



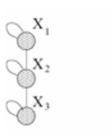
Initial graph



Recursive expansion

  - Does not obey Densification Power Law
  - Has increasing diameter
- Kronecker (tensor) Product is exactly what we need

Kronecker Product – a Graph



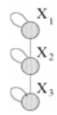
1	1	0
1	1	1
0	1	1

G_1

Adjacency matrix

Kronecker Graphs – Formally:

- We create the self-similar graphs recursively:
  - Start with a initiator graph G_1 on N_1 nodes and E_1 edges
  - The recursion will then product larger graphs $G_2, G_3, \ldots G_k$ on N_1^k nodes
  - Since we want to obey Densification Power Law graph G_k has to have E_1^k edges



G_1

Kronecker Product – Definition

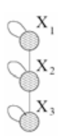
- The Kronecker product of matrices A and B is given by

$$\underset{N \times M}{\mathbf{C} = \mathbf{A}} \otimes \underset{K \times L}{\mathbf{B}} \doteq \underset{N*K \times M*L}{\begin{pmatrix} a_{1,1}\mathbf{B} & a_{1,2}\mathbf{B} & \dots & a_{1,m}\mathbf{B} \\ a_{2,1}\mathbf{B} & a_{2,2}\mathbf{B} & \dots & a_{2,m}\mathbf{B} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n,1}\mathbf{B} & a_{n,2}\mathbf{B} & \dots & a_{n,m}\mathbf{B} \end{pmatrix}}$$

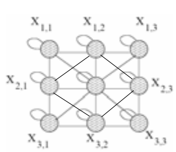
- We define a Kronecker product of two graphs as a Kronecker product of their adjacency matrices

Kronecker Graphs – Intuition

- Intuition:
  - Recursive growth of graph communities
  - Nodes get expanded to micro communities
  - Nodes in sub-community link among themselves and to nodes from different communities



G_1



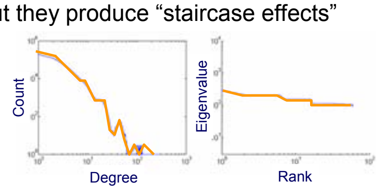
$G_2 = G_1 \otimes G_1$

We can prove Kronecker is realistic

- Given a growing graph with nodes N_1, N_2, …
- We can prove that Kronecker Graphs have the following properties found in real graphs:
  - Static Patterns
    - ✓ Power Law Degree Distribution
    - ✓ Power Law eigenvalue and eigenvector distribution
    - ✓ Small Diameter
  - Temporal Patterns
    - ✓ Densification Power Law
    - ✓ Shrinking/Stabilizing Diameters

Kronecker Graphs

- Kronecker Graphs have all desired properties
- But they produce "staircase effects"



- We introduce a probabilistic version
  Stochastic Kronecker Graphs

How to randomize a graph?

- We want a randomized version of Kronecker Graphs
- Obvious solution
  - Randomly add/remove some edges
- Wrong! – is not biased
  - adding random edges destroys degree distribution, diameter, …
- Want add/delete edges in a biased way
- How to randomize properly and maintain all the properties?

Stochastic Kronecker Graphs

- Create $N_1 \times N_1$ probability matrix P_1
- Compute the k^{th} Kronecker power P_k
- For each entry p_{uv} of P_k include an edge (u,v) with probability p_{uv}

0.5	0.2
0.1	0.3

P_1

Kronecker multiplication

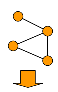
0.25	0.10	0.10	0.04
0.05	0.15	0.02	0.06
0.05	0.02	0.15	0.06
0.01	0.03	0.03	0.09

P_k

Instance Matrix G_2

flip biased coins

Fitting Kronecker to Real Data



- Given a graph G and Kronecker matrix P_1 we can calculate probability that P_1 generated G: $P(G|P_1)$:

0.5	0.2
0.1	0.3

P_1

0.25	0.10	0.10	0.04
0.05	0.15	0.02	0.06
0.05	0.02	0.15	0.06
0.01	0.03	0.03	0.09

P_k

1	1	0	0
1	1	1	0
0	1	1	1
0	0	1	1

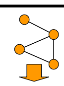
G

$P(G|P_1)$

$$P(G|\mathcal{P},\sigma) = \prod_{(u,v)\in G} \mathcal{P}[\sigma_u,\sigma_v] \prod_{(u,v)\notin G} (1-\mathcal{P}[\sigma_u,\sigma_v])$$

σ… node labeling

Fitting Kronecker: 2 challenges



0.5	0.2
0.1	0.3

P_1

0.25	0.10	0.10	0.04
0.05	0.15	0.02	0.06
0.05	0.02	0.15	0.06
0.01	0.03	0.03	0.09

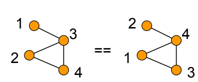
P_k

1	1	0	0
1	1	1	0
0	1	1	1
0	0	1	1

G

$P(G|P_1)$

- Invariance to node labeling (there are N! labelings)
- Calculating P(G|P1) takes $O(N^2)$ (since one needs to consider every cell of adjacency matrix)



Fitting Kronecker: Solutions

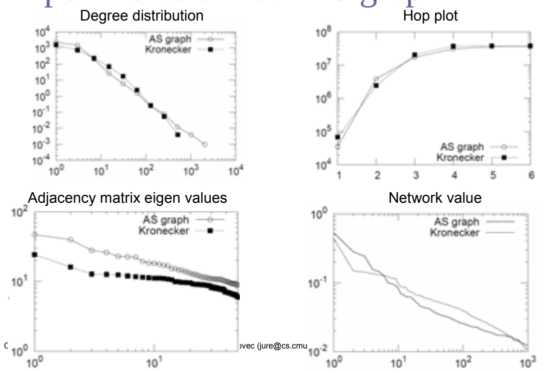
$$P(G|\mathcal{P},\sigma) = \prod_{(u,v)\in G} \mathcal{P}[\sigma_u,\sigma_v] \prod_{(u,v)\notin G} (1-\mathcal{P}[\sigma_u,\sigma_v])$$

σ… node labeling

$$P=\begin{array}{|c|c|c|c|}\hline 0.25 & 0.10 & 0.10 & 0.04 \\ \hline 0.05 & 0.15 & 0.02 & 0.06 \\ \hline 0.05 & 0.02 & 0.15 & 0.06 \\ \hline 0.01 & 0.03 & 0.03 & 0.09 \\ \hline \end{array} \qquad G=\begin{array}{|c|c|c|c|}\hline 1 & 1 & 0 & 0 \\ \hline 1 & 1 & 1 & 0 \\ \hline 0 & 1 & 1 & 1 \\ \hline 0 & 0 & 1 & 1 \\ \hline \end{array}$$

- Node Labeling: can use MCMC sampling (think of it as simulated annealing) to discover good labelings
- $P(G|P_1)$ takes $O(N^2)$: Real graphs are sparse, so calculate $P(G_{empty})$ and then "add" edges. This takes $O(E)$.

Experiments on real AS graph



Why fitting graph modes?

- Parameters tell us about the structure of a graph
- *Extrapolation*: given a graph today, how will it look in a year?
- *Sampling*: can I get a smaller graph with similar properties?
- *Anonymization*: instead of releasing real graph (e.g., email network), we can release a synthetic version of it

Conclusion (1)

- We study evolution of graphs over time
- We discover:
  - Densification Power Law
  - Shrinking Diameters
- Propose explanation:
  - Community Guided Attachment leads to Densification Power Law

$$E(t) \propto N(t)^a$$

CMU 15-826 Jure Leskovec (jure@cs.cmu.edu) 55

Conclusion (2)

- We propose a family of Kronecker Graph generators
- We use the Kronecker Product
- We introduce a randomized version Stochastic Kronecker Graphs
- We fit Kronecker graphs to real data
- And show Kronecker generates graphs with properties similar to those found in real graphs

CMU 15-826 Jure Leskovec (jure@cs.cmu.edu) 56

- References:

  - J. Leskovec, J. Kleinberg, C. Faloutsos: *Graphs over Time: Densification Laws, Shrinking Diameters and Possible Explanations*, In KDD 2005
  - J. Leskovec, D. Chakrabarti, J. Kleinberg, C. Faloutsos: *Realistic, Mathematically Tractable Graph Generation and Evolution, Using Kronecker Multiplication*, In ECML/PKDD 2005
  - Jure Leskovec, Jon Kleinberg, Christos Faloutsos: Graph Evolution: Densification and Shrinking Diameters. ACM Transactions on Knowledge Discovery from Data (ACM TKDD), 1(1), 2007

CMU 15-826 Jure Leskovec (jure@cs.cmu.edu) 57

Propagation of information and influence in networks

Joint work with:
Lada Adamic, University of Michigan
Bernardo Huberman, HP Labs
Natalie Glance and Matthew Hurst, Nielsen Buzzmetrics
Mary McGlohon and Christos Faloutsos

Using online networks for viral marketing

Burger King’s subservient chicken

Motivation for viral marketing

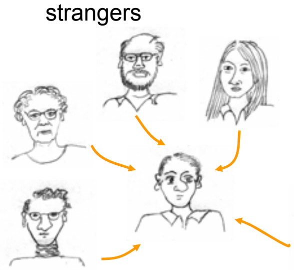
- viral marketing successfully utilizes social networks for adoption of some services
  - hotmail gains 18 million users in 12 months, spending only $50,000 on traditional advertising
  - gmail rapidly gains users although referrals are the only way to sign up
- customers becoming less susceptible to mass marketing
- mass marketing impractical for unprecedented variety of products online
  - Google AdSense helps sellers reach buyers with targeted advertising
  - but how do buyers get good recommendations?

The web savvy consumer and personalized recommendations

- > 50% of people do research online before purchasing electronics
- personalized recommendations based on prior purchase patterns and ratings
  - Amazon, “people who bought x also bought y”
  - MovieLens, “based on ratings of users like you…”
- Is there still room for viral marketing?

next to personalized recommendations?

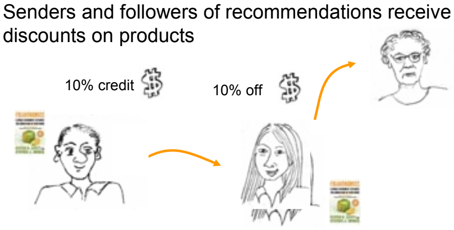
- We are more influenced by our friends than strangers



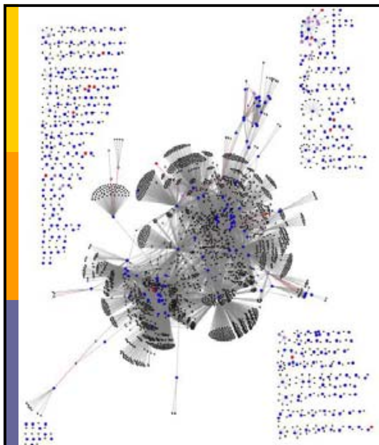

- 68% of consumers consult friends and family before purchasing home electronics (Burke 2003)

Incentivised viral marketing (our problem setting)

- 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

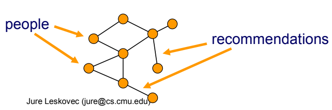
the recommendation network data

- large anonymous online retailer (June 2001 to May 2003)
- 15,646,121 recommendations
- 3,943,084 distinct customers
- 548,523 products recommended
- Products belonging to 4 product groups:
  - books
  - DVDs
  - music
  - VHS

summary statistics by product group

	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

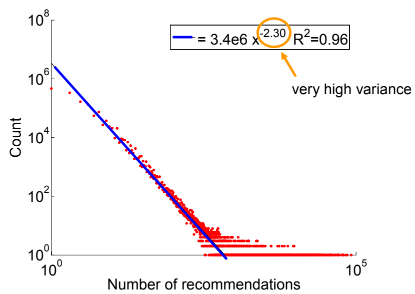


viral marketing program not spreading virally

- 94% of users make first recommendation without having received one previously
- 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

CMU 15-826 Jure Leskovec (jure@cs.cmu.edu) 67

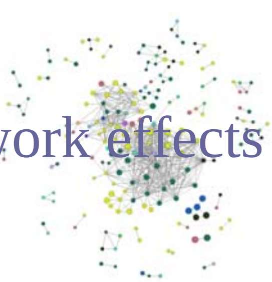
participation level by individual



The most active person made 83,729 recommendations and purchased 4,416 different items!

CMU 15-826 68

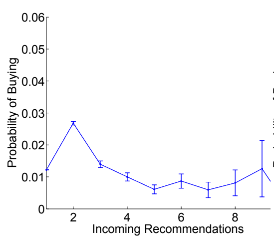
Network effects



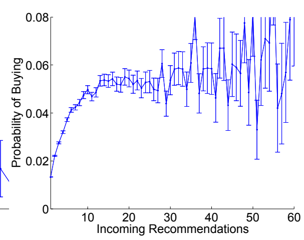
CMU 15-826 Jure Leskovec (jure@cs.cmu.edu) 69

does receiving more recommendations increase the likelihood of buying?

BOOKS

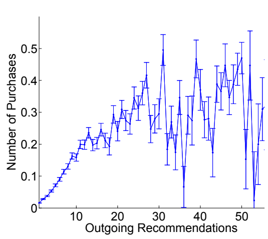


DVDs

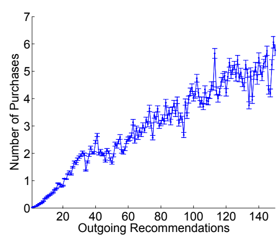


does sending more recommendations influence more purchases?

BOOKS

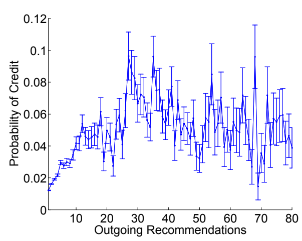


DVDs



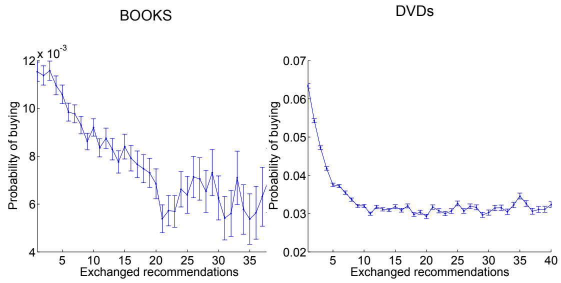
the probability that the sender gets a credit with increasing numbers of recommendations

- 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

Multiple recommendations between two individuals weaken the impact of the bond on purchases



CMU 15-826 Jure Leskovec (jure@cs.cmu.edu) 73

product and social network characteristics influencing recommendation effectiveness

CMU 15-826 Jure Leskovec (jure@cs.cmu.edu) 74

recommendation 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 (significant at p=0.01 level)
  - 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 (statistically significant)
  - 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)

CMU 15-826 Jure Leskovec (jure@cs.cmu.edu) 75

anime DVDs

- 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

CMU 15-826 Jure Leskovec (jure@cs.cmu.edu) 76

regressing on product characteristics

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

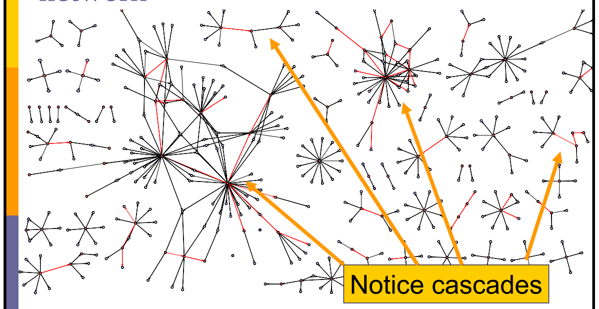
CMU 15-826 Jure Leskovec (jure@cs.cmu.edu) 77

products most suited to viral marketing

- small and tightly knit community
  - few reviews, senders, and recipients
  - but sending more recommendations helps
- pricey products
- rating doesn't play as much of a role

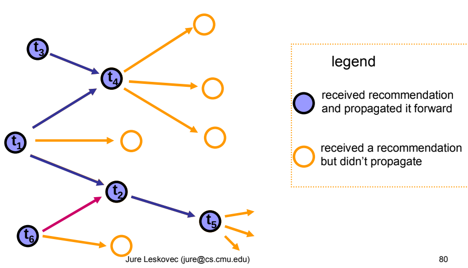
CMU 15-826 Jure Leskovec (jure@cs.cmu.edu) 78

medical study guide recommendation network



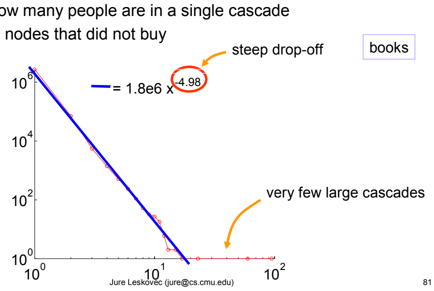
Cascade formation process

- Time: $t_1 < t_2 < \ldots < t_n$



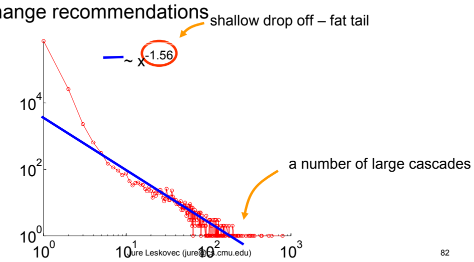
measuring cascade sizes

- delete late recommendations
- count how many people are in a single cascade
- exclude nodes that did not buy



cascades for DVDs

- DVD cascades can grow large
- possibly as a result of websites where people sign up to exchange recommendations



CMU 15-826 Jure Leskovec (jure@cs.cmu.edu) 82

simple model of propagating recommendations

(ignoring for the moment the specific mechanics of the recommendation program of the retailer)

- Each individual will have p_t successful recommendations. We model p_t as a random variable.
- At time $t+1$, the total number of people in the cascade,

$$N_{t+1} = N_t * (1+p_t)$$

- Subtracting from both sides, and dividing by N_t, we have

$$\frac{N_{(t+1)} - N_t}{N_t} = p_t - 1$$

CMU 15-826 Jure Leskovec (jure@cs.cmu.edu) 83

simple model of propagating recommendations

(continued)

- Summing over long time periods

$$\frac{dN}{N} = \sum p_t$$

- The right hand side is a sum of random variables and hence normally distributed.
- Integrating both sides, we find that N is lognormally distributed

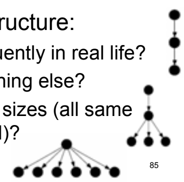
$$P(N) = \frac{1}{N\sqrt{2\pi\sigma^2}} \exp\frac{-(\ln(N)-\mu)^2}{2\sigma^2}$$

if σ large resembles power-law

CMU 15-826 Jure Leskovec (jure@cs.cmu.edu) 84

Cascade shapes

- We look at the fine-grained patterns of influence in a large-scale, real recommendation network
- Given a directed who-influences-whom graph
- Find cascades
- And examine their topological structure:
  - What kinds of cascades arise frequently in real life?
  - Are they like trees, stars, or something else?
  - What is the distribution of cascade sizes (all same size / exponential tail / heavy-tailed)?

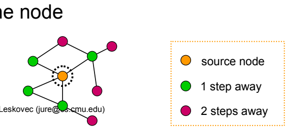


Identifying cascades

- Given a set of recommendations find cascades
- We use the following approach
  - Create a separate graph for each product
  - Delete late recommendations:
    - Delete recommendations that happened after the first purchase of the product
    - We get time-increasing graph
  - Delete no-purchase nodes:
    - We find many star-like patterns, no propagation of influence
    - Delete nodes that did not purchase a product
  - Now connected components correspond to maximal cascades

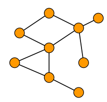
Cascade enumeration

- Maximal cascades do not reveal what are the cascade building blocks (local structures)
- Given a maximal cascade we want to enumerate all local cascades:
  - For every node we explore the cascade in the neighborhood up to 1, 2, 3,… steps away
  - This way we capture the local structure of the cascade around the node

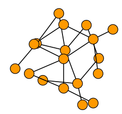


Counting cascades (graph isomorphism)

- To count cascades we need to determine whether a new cascade is isomorphic to already seen one:



? ==



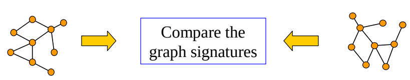
Graphs are isomorphic if there exists a node mapping so that nodes have same neighbors

- No polynomial graph isomorphism algorithm is known, so we reside to approximate solution

CMU 15-826 Jure Leskovec (jure@cs.cmu.edu) 88

Graph isomorphism

- Do not compare the graphs directly, but
- For each graph we create a signature
- A good signature is one where isomorphic graphs have the same signature, but few non-isomorphic graphs share the same signature



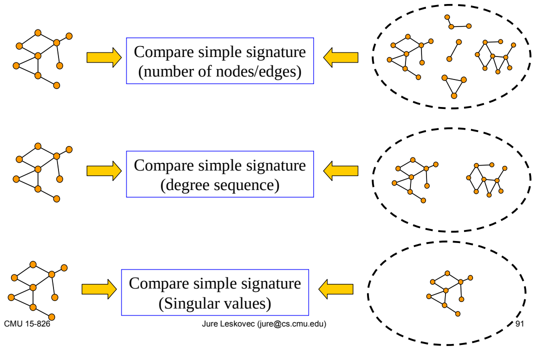
CMU 15-826 Jure Leskovec (jure@cs.cmu.edu) 89

Comparing signatures

- First compare simple signatures
- Compare the graphs with the same simple signature using more and more complicated (expensive/accurate) signatures
- At the end (for small graphs) we perform exact isomorphism resolution
- Since we are interested in building blocks of cascades which are generally small, the precision for small graphs is more important

CMU 15-826 Jure Leskovec (jure@cs.cmu.edu) 90

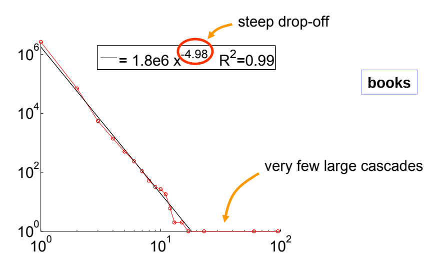
Comparing signatures – Example



CMU 15-826 Jure Leskovec (jure@cs.cmu.edu) 91

Measuring maximal cascade sizes

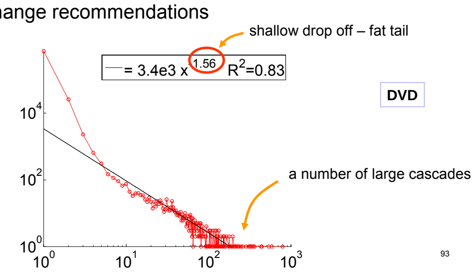
- Count how many people are in a single cascade
- We observe a heavy tailed distribution which can not be explained by a simple branching process



CMU 15-826 92

Cascade sizes for DVDs

- DVD cascades can grow large
- possibly a product of websites where people sign up to exchange recommendations



CMU 15-826 93

Frequent cascade subgraphs (1)

high
low

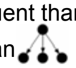
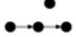
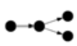
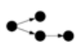
- General observations:
  - DVDs have the richest cascades (most recommendations, most densely linked)
  - Books have small cascades
  - Music is 3 times larger than video but does not have much variety in cascades

	cascades	different
Book	122,657	959
DVD	289,055	87,614
Music	13,330	158
Video	1,928	109
	number of all "words"	vocabulary size

CMU 15-826 Jure Leskovec (jure@cs.cmu.edu) 94

Frequent cascade subgraphs (2)

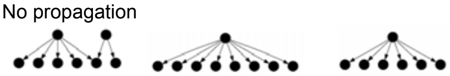
 is the most common cascade subgraph
- It accounts for ~75% cascades in books, CD and VHS, only 12% of DVD cascades

 is 6 (1.2 for DVD) times more frequent than 
- For DVDs  is more frequent than 
- Chains () are more frequent than 
-  is more frequent than a collision () (but collision has less edges)
- Late split () is more frequent than 

CMU 15-826 Jure Leskovec (jure@cs.cmu.edu) 95

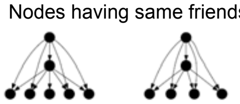
Typical classes of cascades

- No propagation

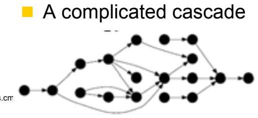


- Common friends

- Nodes having same friends

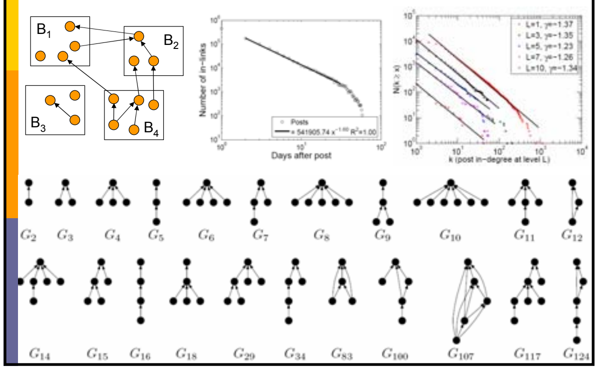


- A complicated cascade



CMU 15-826 Jure Leskovec (jure@cs.cm

Propagation of information in Blogs



Conclusions

Overall

- incentivized viral marketing contributes marginally to total sales
- occasionally large cascades occur

Observations for future 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

Conclusions (2)

- Cascades are a form of collective behavior
- From our experiments we found:
  - Most cascades are small, but large bursts can occur
  - Cascade sizes follow a heavy-tailed distribution
  - Frequency of different cascade subgraphs depends on the product type
  - Cascade frequencies do not simply decrease monotonically for denser subgraphs
  - But reflect more subtle features of the domain in which the recommendations are operating

- References:
  - The Dynamics of Viral Marketing. Jure Leskovec, Lada Adamic, Bernardo Huberman. ACM Conference on Electronic Commerce (EC 2006)
  - Patterns of Influence in a Recommendation Network. Jure Leskovec, Ajit Singh, Jon Kleinberg. Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2006)
  - Cascading Behavior in Large Blog Graphs. Jure Leskovec, Mary McGlohon, Christos Faloutsos, Natalie Glance, Matthew Hurst. SIAM International Conference on Data Mining (SDM 2007).