IIIT-H

Web Mining
Lecture 10: Social Network Analysis (Part 2)

Manish Gupta

2^{nd} Sep 2013

Slides borrowed (and modified) from
http://www.stanford.edu/class/cs224w/slides/11-powerlaws.pdf
http://temporalweb.net/2011/files/kumar-twaw2011.pdf
http://www.stanford.edu/class/cs224w/slides/12-evolution.pdf

Recap of Lecture 9: Social Network Analysis (Part 1)

- Introduction to Social Network Analysis
- Structure of the Web Graph
- Erdös-Renyi Model
- Small World Model and Kleinberg's Model
- Power Laws

Announcements

- Midsem Exam: Sep 6, 1:30pm-3pm
- Assignment 2
  - Submission Deadline is Sep 5, 9pm
  - Any form of copying will lead to 0 marks for everyone with the same answer
- Midsems
  - No cheating
  - Allowed 1 A4 size cheat sheet (both sides)
  - Same format as Assignment 2
  - Covers content covered up to Aug 22 (last lecture)
- Doubt clearing session
  - Sep 2, 7:30pm-9pm
  - TAs will conduct this one at 104 Himalaya

Today's Agenda

- Preferential Attachment Model
- Copying Model, Forest Fire Model
- Model with Network Components
- Evolving Network Model
- Compressible Graph Model

Today's Agenda

- **Preferential Attachment Model**
- Copying Model, Forest Fire Model
- Model with Network Components
- Evolving Network Model
- Compressible Graph Model

Preferential Attachment Model (1)

- [Price ‘65, Albert-Barabasi ’99, Mitzenmacher ‘03]
- Nodes arrive in order 1,2,…,n
- At step j, let d_i be the degree of node $i < j$
- A new node j arrives and creates m outlinks
- Probability of j linking to previous node i is proportional to the degree d_i of node i. $P(j \longrightarrow i) = \frac{d_i}{\sum_k d_k}$
- Rich get richer
  - New nodes are more likely to link to nodes that already have high degree
  - Herbert Simon’s result: Power-laws arise from “Rich get richer” (cumulative advantage)
  - Examples[Price 65]
    - Citations: New citations to a paper are proportional to the number it already has

Preferential Attachment Model (2)

- Let us analyze the following model
  - Nodes arrive in order 1,2,…, n
  - When node is created it makes a single out-link to an earlier node i chosen
    - With probability p, j links to i chosen uniformly at random (from among all earlier nodes)
    - With probability 1-p, j chooses node i uniformly at random and links to a node i points to
      - That is, with probability 1-p, node j links to node u with probability proportional to d_u (in-degree of node u)
  - Graph is directed; each node has out-degree=1

Preferential Attachment Model (3)

- What is the change in in-degree of a node over time?
  - At time $t > i$
    - t is the number of nodes that have arrived so far
    - Let $d_i(t)$ be the in-degree of node i at time t
    - Initial condition: $d_i(t) = 0$ at $t = i$
    - Expected change in $d_i(t)$ over time
      - Node i gets an in-link at time $t + 1$ only if a link from a newly created node $t + 1$ points to it
        - » With prob p, node $t + 1$ links randomly
          - Prob of linkage to node i is $\frac{1}{t}$
        - » With prob 1-p, node $t + 1$ links preferentially
          - Prob of linkage to node i is $\frac{d_i(t)}{t}$
      - Prob. that node $t + 1$ links to node i is $p\frac{1}{t} + (1 - p)\frac{d_i(t)}{t}$

Preferential Attachment Model (4)

- What is the rate of growth of $d_i(t)$?
  - $\frac{dd_i(t)}{dt} = p\frac{1}{t} + (1-p)\frac{d_i(t)}{t} = \frac{\mathrm{p+qd_i(t)}}{\mathrm{t}}$
  - $\int \frac{1}{p+qd_i(t)} dd_i(t) = \int \frac{1}{t} dt$
  - $\frac{1}{q}\ln\left(p + qd_i(t)\right) = \ln t + c$
  - $d_i(t) = \frac{1}{q}(At^q - p)$
  - Now $d_i(i) = 0 \Rightarrow A = \frac{p}{i^q}$
  - Hence, $d_i(t) = \frac{p}{q}\left(\left(\frac{t}{i}\right)^q - 1\right)$

Preferential Attachment Model (5)

- What is the degree distribution for this model?
  - At time t, what is $F(k)$, the fraction of nodes with degree at least k?
    - That is, how many nodes have degree>k
      - $d_i(t) = \frac{p}{q}\left(\left(\frac{t}{i}\right)^q - 1\right)$>k
      - Solving for i gives $\mathrm{i} < \mathrm{t}\left(\frac{\mathrm{q}}{\mathrm{p}}\mathrm{k} + 1\right)^{-\frac{1}{\mathrm{q}}}$
    - Since there are t nodes at time t, fraction $F(k) = \left(\frac{\mathrm{q}}{\mathrm{p}}\mathrm{k} + 1\right)^{-\frac{1}{\mathrm{q}}}$
  - At time t, what is the fraction of nodes with degree exactly k?
    - $F(k)$ is CDF, so $F'(k)$ will be the PDF
    - $F'(k) = \frac{1}{p}\left(\frac{\mathrm{q}}{\mathrm{p}}\mathrm{k} + 1\right)^{-1-\frac{1}{\mathrm{q}}}$
    - Thus, the degree distribution is a power law with exponent $\alpha = 1 + \frac{1}{1-p}$
    - For web, $\alpha = 2.1 \Rightarrow p \sim 0.1$

Preferential Attachment Model (6)

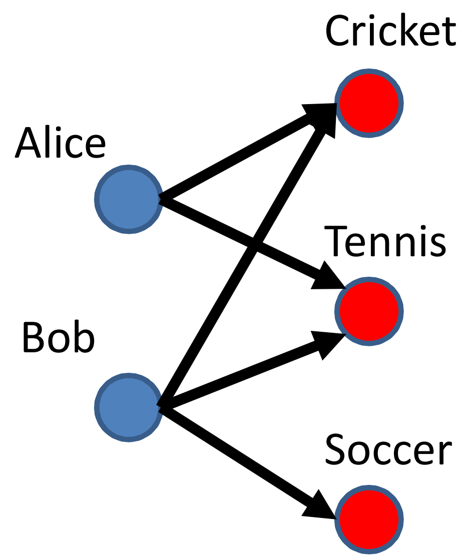
- Other network formation mechanisms that generate scale-free networks
  - Random surfer model [Blum-Mugizi]
  - Copying model [Kleinberg et al.]
  - Forest Fire model [Leskovec et al.]

Today's Agenda

- Preferential Attachment Model
- **Copying Model, Forest Fire Model**
- Model with Network Components
- Evolving Network Model
- Compressible Graph Model

Communities in Social Networks

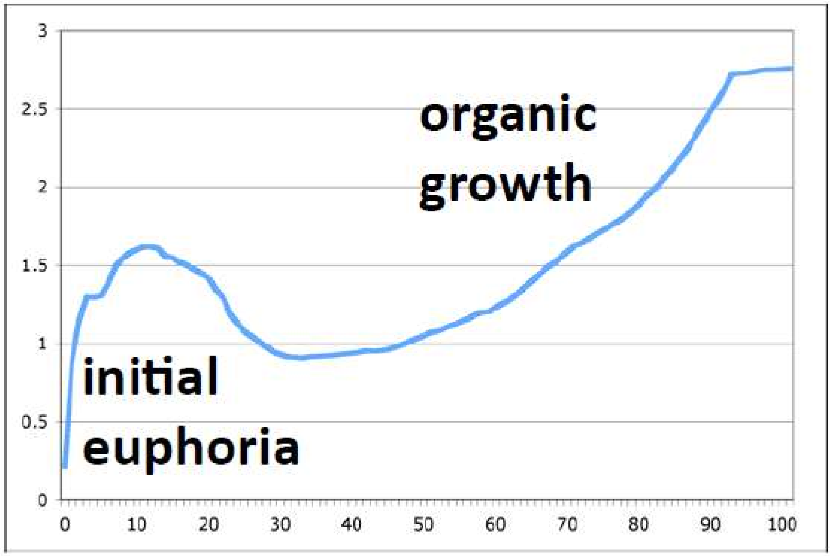
- Edges usually imply endorsement or interest in a topic or a person
  - Users link to pages they care about
    - Two users with similar interest need not know each other
  - Friendship links in social networks
- Communities are dense subgraphs or dense bipartite subgraphs
- Web and social networks are abundant in communities



Copying Model [Kumar et al., 2000]

- Observation: People copy their friend's webpage when creating a new one or copy their friend's contacts when joining a social network
- When a new node arrives, it copies edges from a pre-existing node with probability 1 – a links to the destination of the edge
- The degree distribution ¡s a power-law with exponent $\frac{2-\alpha}{1-\alpha}$
- Can explain communities: The number of dense bipartite cliques in this model is large

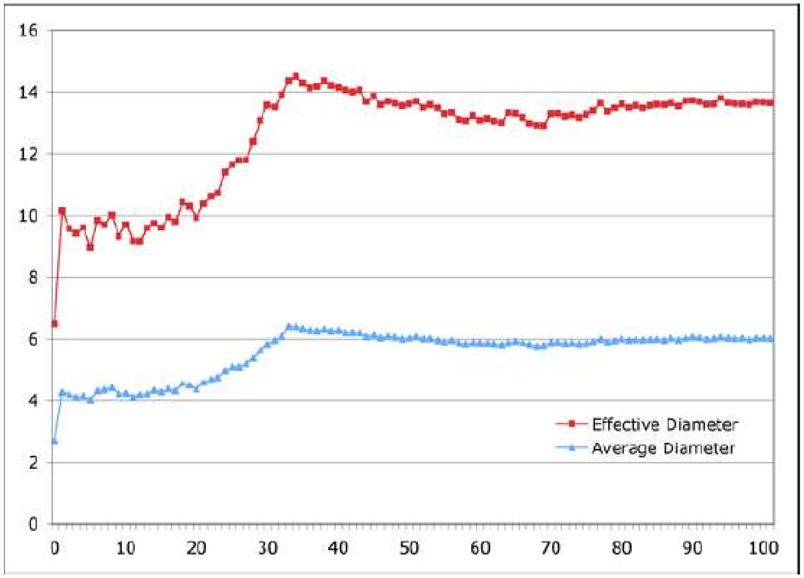
Flickr: Density, Diameter over Time



Density increases over time

Shrinking diameters and densification in citation graphs: **Leskovec, Kleinberg, Faloutsos** **2005**

Diameter shrinks over time



Forest Fire Model [Leskovec, Kleinberg, Faloutsos 2005]

- Observation: Copying happens beyond one step
- When a new node arrives, it
  - copies an edge from a pre-existing node with prob. 1 - a
  - copies an edge from the destination of the edge
  - …
- An iterated version of the copying model
- In addition to the above, leads to densification and shrinking diameters, in empirical simulations

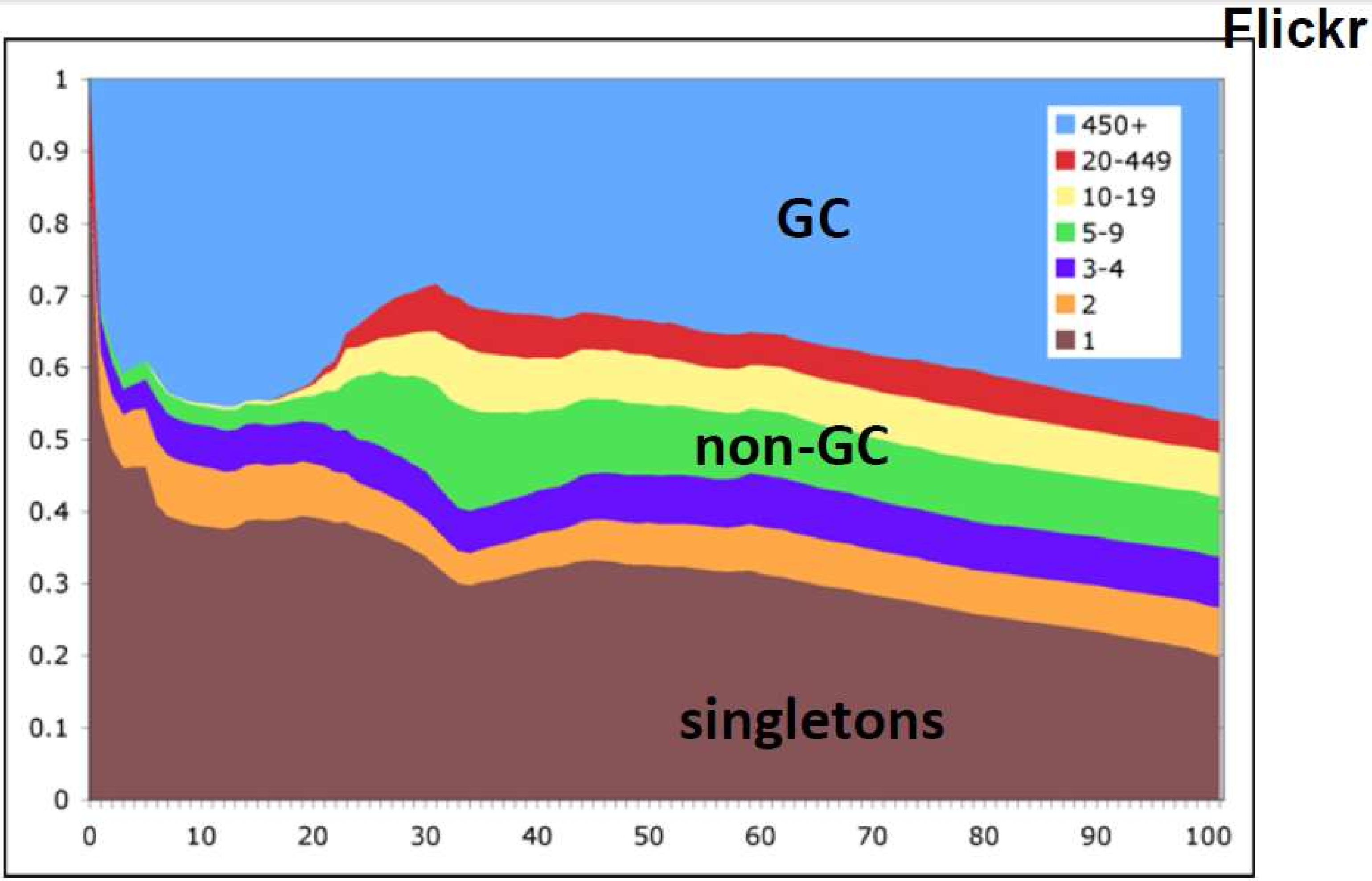
Affiliation Networks Model

- Bipartite graph B(Q, U, D) and a graph G(Q, E)
  - Q = papers, U = topics
- Co-evolution of B and G
  - Q side of B evolves by copying
  - U side of B evolves by copying
  - Q side of G evolves by prototyping (via evolution of Q and U in B)
- This evolutionary model produces graphs with densification and shrinking diameters [Lattanzi, Sivakumar 2009]

Today's Agenda

- Preferential Attachment Model
- Copying Model, Forest Fire Model
- **Model with Network Components**
- Evolving Network Model
- Compressible Graph Model

Components: A Grand Canyon View

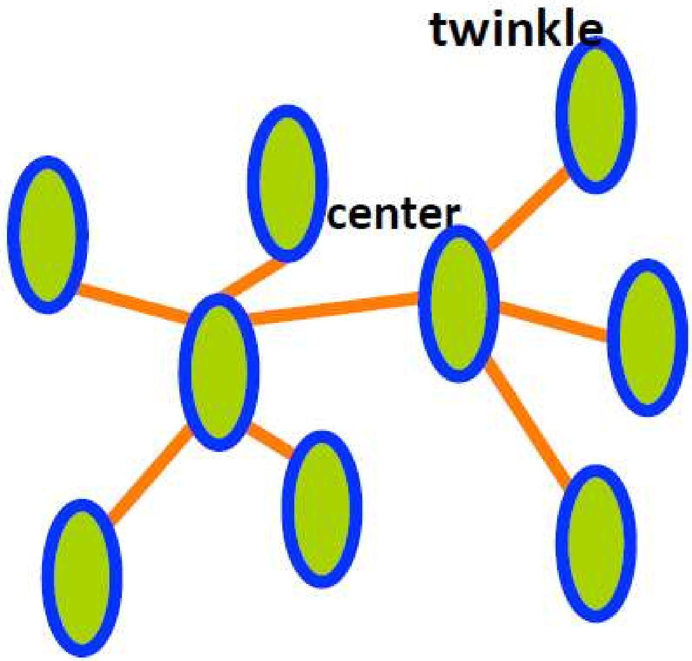


How do Components Consolidate?

- Singletons merge with non-GC and GC
- Non-GCs merge with GC
- Almost never a non-GC merges with another non-GC
- Why is singleton attracted to a non-GC?
  - Is there a special attractor in a non-GC?

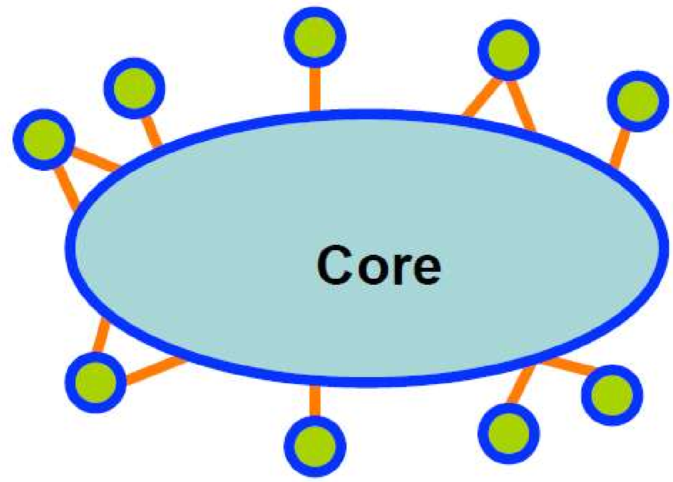
Structure of Non-GCs: Stars

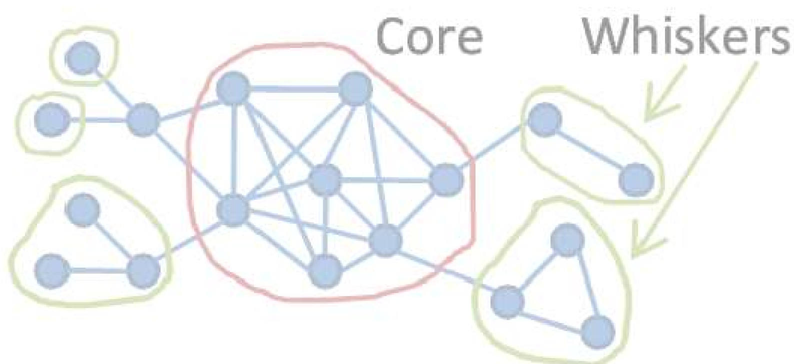
- There are one or more centers (high degree nodes)
- There are many degree-1 nodes (twinkles) connected to these centers
- Under reasonable setting of parameters, around 93% of non-GCs are stars
- The stars form quickly
- A large fraction of them are yet to be absorbed into GC



Structure of the GC: Core

- There is a small core of very high connectivity inside GC
- The core is not comprised of star centers
- GC connectivity does not depend on star centers
- This has implications for finding dense communities: Leskovec, Lang, Dasgupta, Mahoney, 2008





A Simple Model with User Types

- At each time step
  - A person joins the network and is chosen to be one of three types: passive user, inviter, linker
  - Few friendships (i.e., edges) arrive
    - Source of edge chosen from inviters/linkers with degree-biased prob (i.e., preferential attachment)
    - If source=inviter, destination=a new passive user
    - If source=linker, destination chosen from linkers and inviters, degree-biased
- Empirically, this model generates the observed temporal characteristics (fraction of components, stars, core)

Today's Agenda

- Preferential Attachment Model
- Copying Model, Forest Fire Model
- Model with Network Components
- **Evolving Network Model**
- Compressible Graph Model

Comparing Models

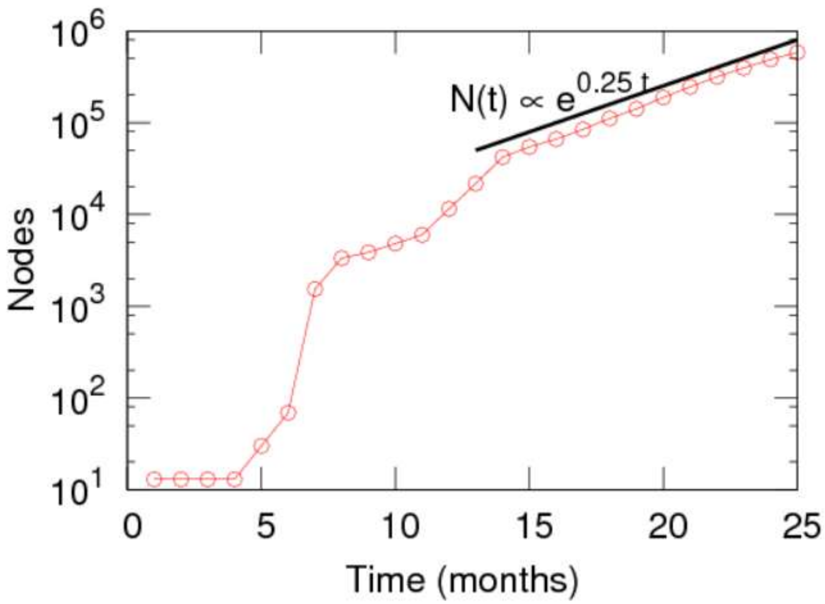
- We have so many models
- What is the best way to compare two graph models?
  - Maximum-likelihood (standard tool in ML)
  - Efficiency issues: Bezakova, Kalai, Santhanam, 2006
- We have edge-by-edge arrival information, so can take an edge and compare the likelihood of its existence in competing models

How does the Network Evolve?

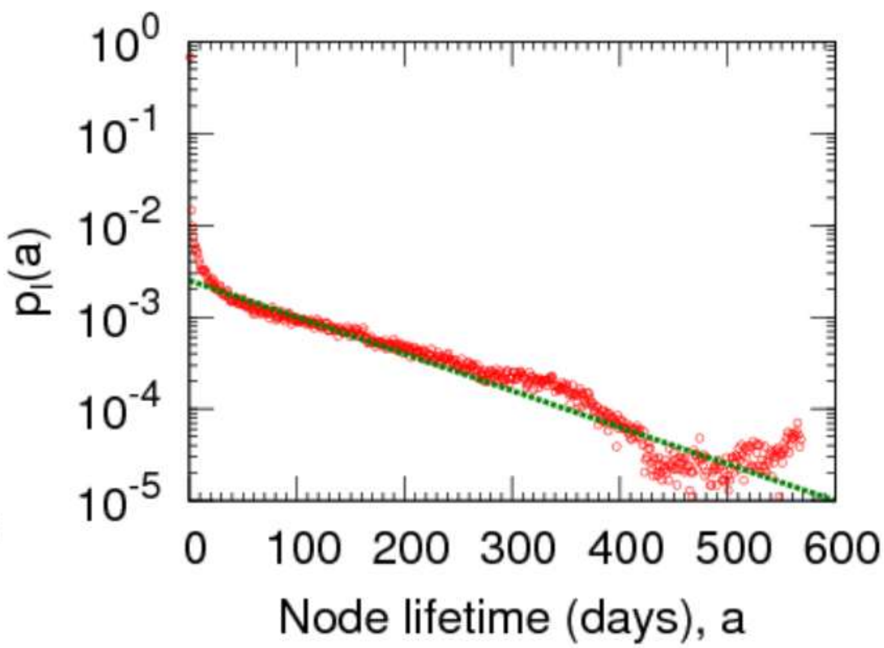
- Three processes govern the evolution
  - Node arrival process: Nodes enter the network
  - Edge initiation process: Each node decides when to initiate an edge
  - Edge destination process: Determines destination after a node decides to initiate
- We will present a complete model of network evolution by mining the node and edge creation data

Node Arrivals: Rate and Lifetime

- Node arrival rate really depends on the network, ranging from sub-linear to exponential
- Lifetime a: time between node's first and last edge
  - Node lifetime is exponentially distributed: $p_l(a) = \lambda e^{-\lambda a}$



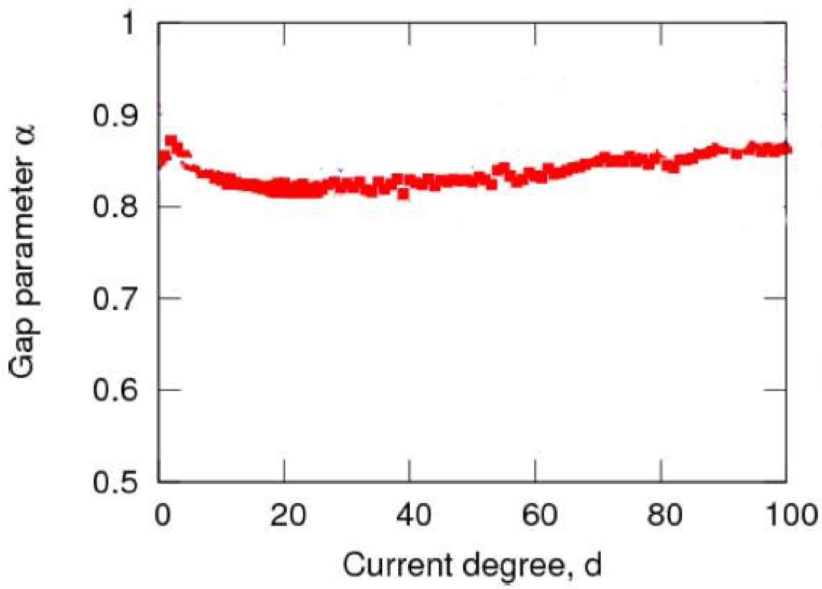
Node Arrival Rate

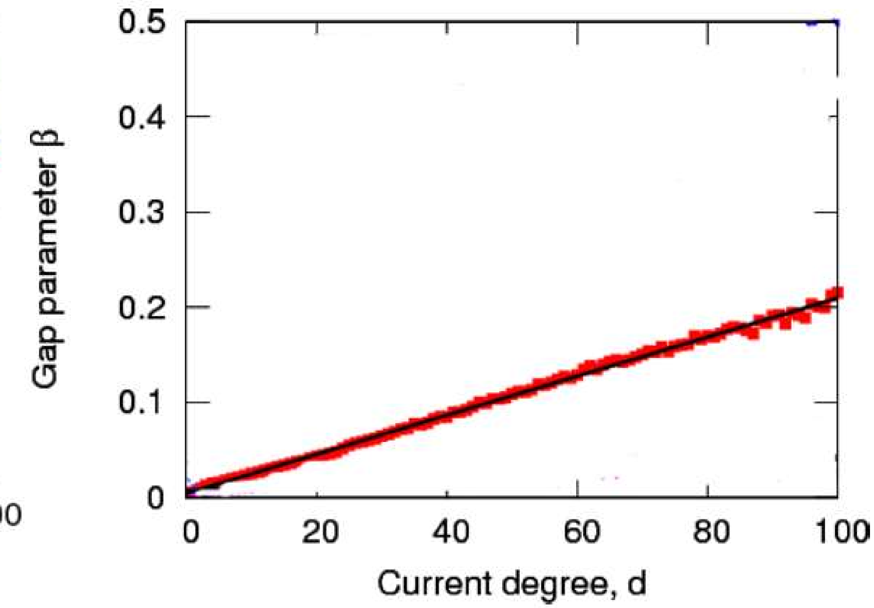


How are the Edges Initiated?

- Let $\delta(d)$ be the edge gap, i.e., the time between d^{th} and $d+1^{st}$ edge
- Competing models: exponential, log normal, stretched exponential, power-law with exponential cutoff
- Edge inter-arrivals follow power-law with exponential cutoff: $p_g\big(\delta(d); \alpha(d), \beta(d)\big) \propto \delta(d)^{-\alpha(d)} e^{-\beta(d)\delta(d)}$

How do α and β Change with Degree?

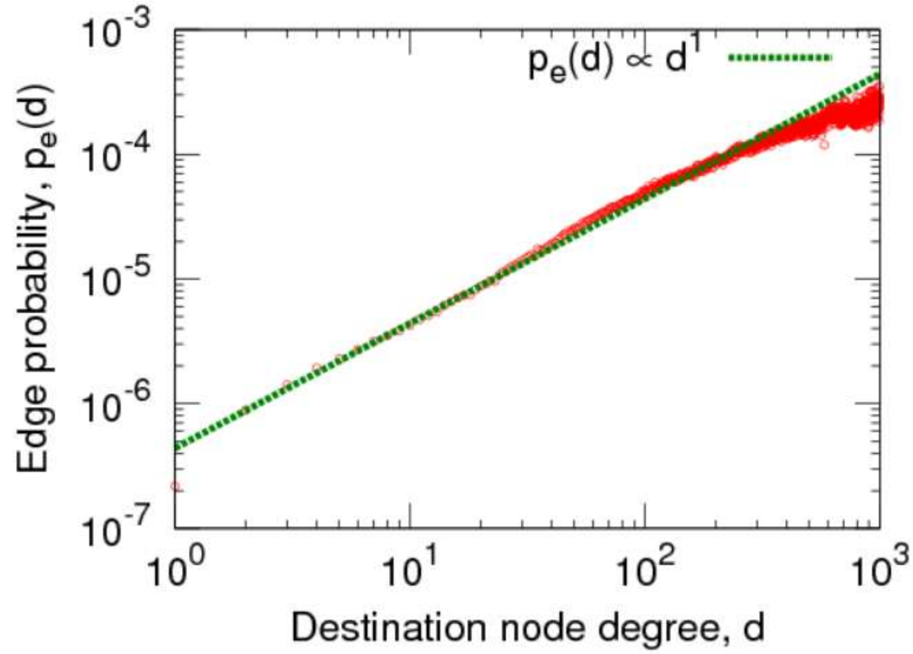




- $\alpha(d)$ (power law part) is constant
- $\beta(d)$ (exp-cutoff part) is linear in d
- This means nodes of higher degree start adding edges faster and faster
- Next: How to model edge destination?

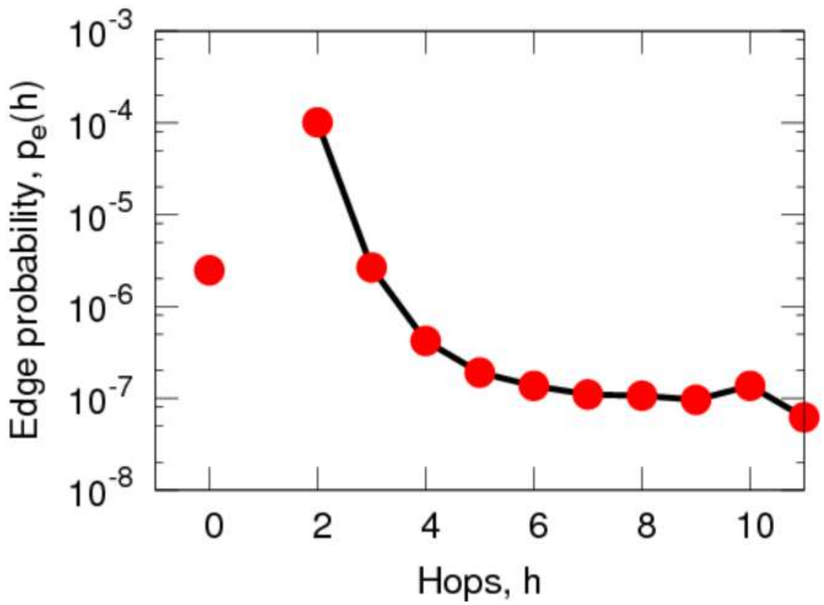
Does Preferential Attachment Happen?

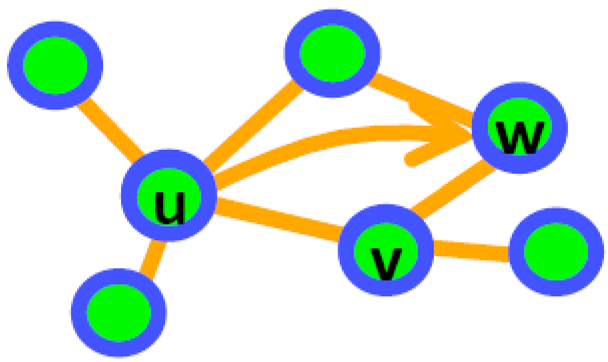
- We unroll the true network edge arrivals and measure node degrees where edges attach
- Preferential attachment indeed happens!
- But there is more to it



How Local are the Added Edges?

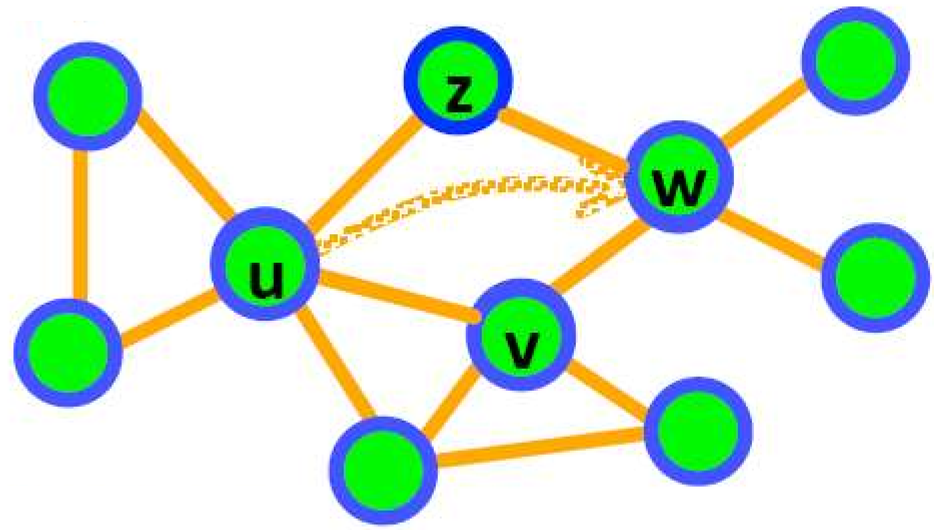
- Just before edge (u,v) is placed, how far are u and v?
- Normalize this by number of nodes at that hop distance
- Real edges are local and most of them (66%) are triangle closing
- Long known to sociologists [George Simmel (1858-1918), Krackhardt and Handcock 2007]





Closing Triangles

- New triangle closing edge (u,w) appears next
- We model this as
  - Choose u's neighbor v
  - Choose v's neighbor w
  - Add edge (u,w)
- 25 strategies for choosing v and then w
  - Random, degree preferentially, number of common friends, time of last activity, combination
- Can compute the likelihood of each strategy



Triangle Closing Strategies

- Log likelihood improvement over the baseline

Strategy to select v (1st node)

Select w (2nd node)

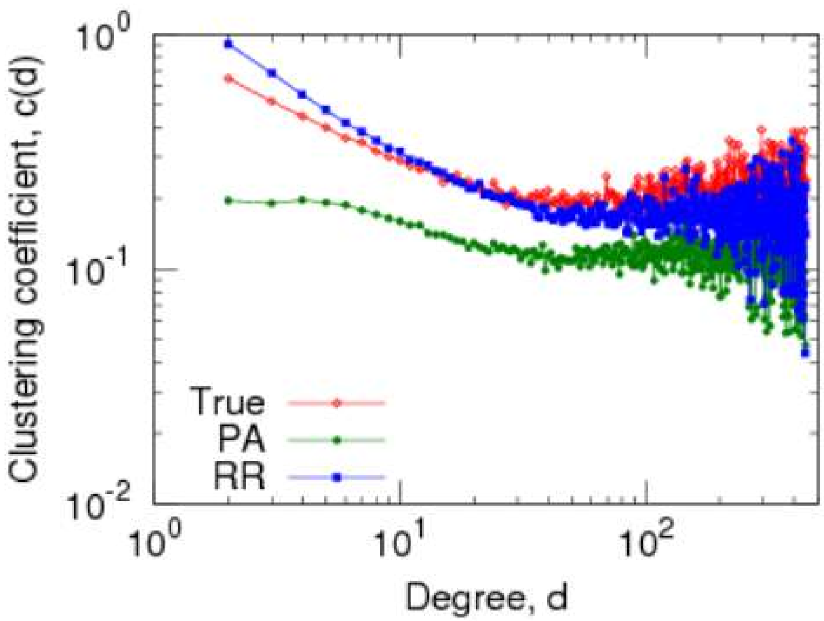
FLICKR	random	$\text{deg}^{0.2}$	com	$\text{last}^{-0.4}$	$\text{comlast}^{-0.4}$
random	13.6	13.9	14.3	16.1	15.7
$\text{deg}^{0.1}$	13.5	14.2	13.7	16.0	15.6
$\text{last}^{0.2}$	14.7	15.6	15.0	17.2	**16.9**
com	11.2	11.6	11.9	13.9	13.4
$\text{comlast}^{0.1}$	11.0	11.4	11.7	13.6	13.2

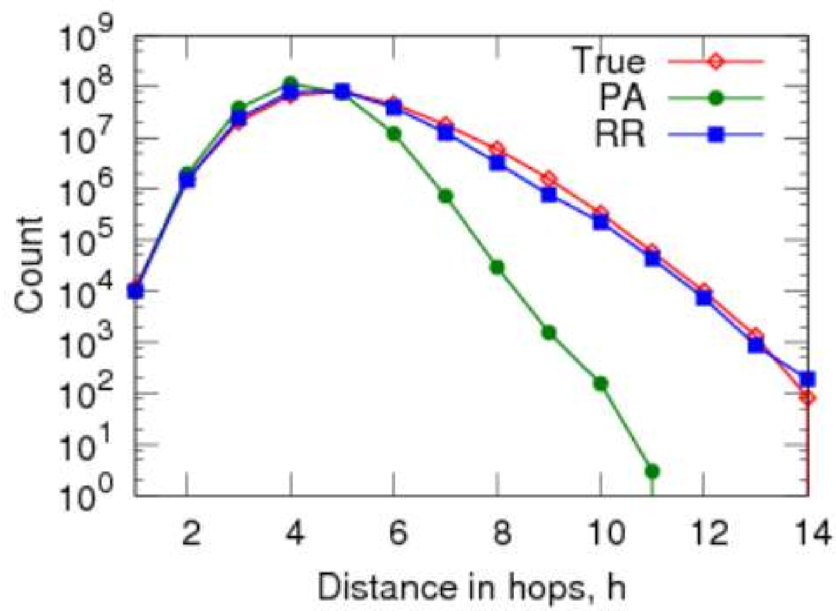
- Strategies to pick a neighbor
  - random: uniformly at random
  - deg: prop. to its degree
  - com: prop. to number of common friends
  - last: prop. to time since last activity
  - comlast: prop. to com*last

Random-random works quite well

Preferential Attachment vs. Random-Random: Semi-Simulation

- Take the network at T/2 and evolve it using preferential attachment (PA) and random-random (RR) for edge addition events





Summary of Evolving Network Generation

- Node arrival
  - Node arrival is network dependent
  - Node lifetime: $p(a) = \lambda e^{-\lambda a}$
- Edge initiation
  - Edge gaps: $p(\delta) \propto \delta^{-\alpha} e^{-\beta d \delta}$
- Edge destination
  - First edge chosen preferentially
  - Use random-random strategy to close triangles

Complete Evolving Network Model

- Nodes arrive using the arrival rate
- Node u arrives
  - It has lifetime a $\sim \lambda e^{-\lambda a}$
  - Adds first edge to node v with prob. proportional to degree of node v
- A node u with degree d has gap $\delta \sim \delta^{-\alpha} e^{-\beta d \delta}$ and goes to sleep for δ time steps
- When u wakes up, if its lifetime is still valid, it creates a random-random triangle-closing edge

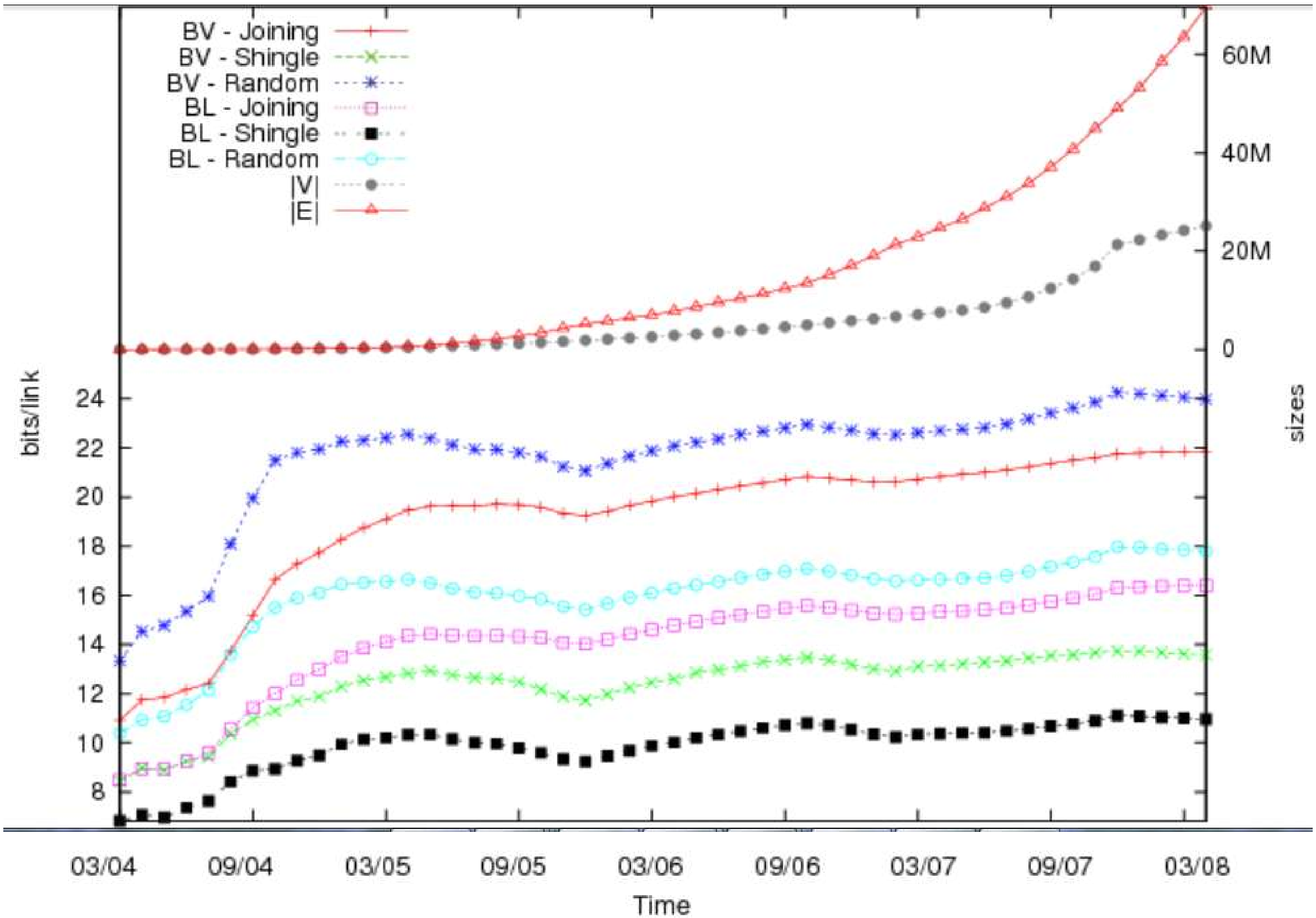
An Analysis of the Evolving Network Model

- The out-degrees are distributed according to a power-law with exponent $1 + \frac{\lambda}{\beta} . \frac{\Gamma(2-\alpha)}{\Gamma(1-\alpha)}$
- For Flickr, true exponent=1.73, λ=0.0092, α=0.84, β=0.002, calculated exponent=1.74
- Analogous results hold for delicious Yahoo! Answers, LinkedIn
- Interesting as temporal behavior leads to power-law degree distribution

Today's Agenda

- Preferential Attachment Model
- Copying Model, Forest Fire Model
- Model with Network Components
- Evolving Network Model
- **Compressible Graph Model**

Compressibility of Flickr over Time



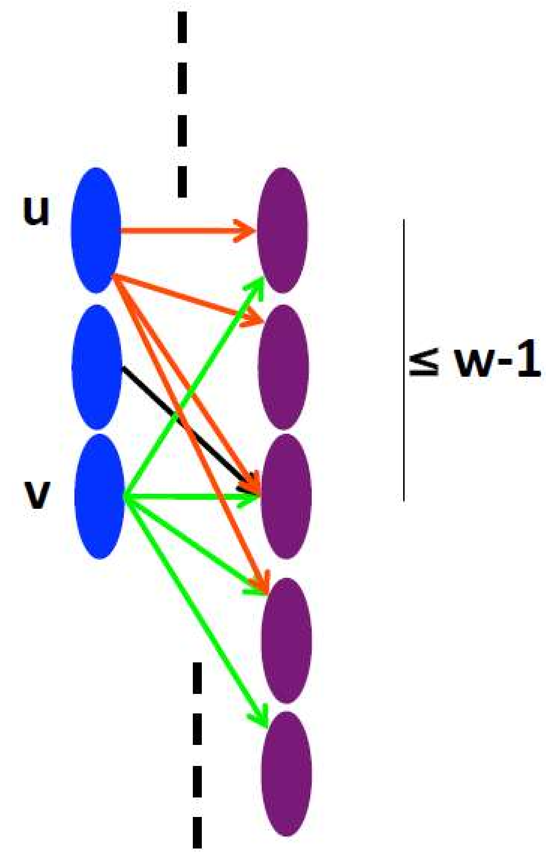
Compressing the Web [Boldi Vigna WWW04]

- Key ideas
  - Many web pages have similar set of neighbors
  - Edges tend to be local
- Canonical Ordering: Sort URLs lexicographically, treating them as strings [Randall et al. 2002]
  - 17: www.uchicago.edu/alchemy
  - 18: www.uchicago.edu/biology
  - 19: www.uchicago.edu/biology/plant
  - 20: www.uchicago.edu/biology/plant/copyright
  - 21: www.uchicago.edu/biology/plant/people
  - 22: www.uchicago.edu/chemistry
- This gives an identifier for each URL

- Source and destination of edges are likely to get nearby IDs
  - Templated webpages
  - Many edges are intra-host or intra-site
- Due to templates, the adjacency list of a node is similar to one of the 7 preceding URLs in the lexicographic ordering
- Express adjacency list in terms of one of these
- E.g., consider the adjacency lists
  - 1,2,4, 8, 16, 32, 64
  - 1, 4, 9, 16, 25, 36, 49, 64
  - 1, 2, 3, 5, 8, 13, 21, 34, 55, 89, 144
  - 4: 1, 4, 8, 16, 25, 36, 49, 64
- Encode list 4 using list 2: remove 9, add 8

BV Compression Algorithm

- Each node has a unique ID from the canonical ordering
- Let w = copying window parameter
- To encode a node v
- Check if out-neighbors of v are similar to any of w-1 previous nodes in the ordering
- If yes, let u be the leader: use log w bits to encode the gap from v to u + difference between out-neighbors of u and v
- If no, write log w zeros and encode out-neighbors of v explicitly
- Use gap encoding on top of this



Canonical Orderings

- BV compressions depend on a canonical ordering of nodes
- This canonical ordering should exploit neighborhood similarity and edge locality
- How do we get a good canonical ordering?
- Unlike the web page case, it is unclear if social networks have a natural canonical ordering
- Caveat: BV is only one genre of compression scheme
- Lack of good canonical ordering does not mean graph is incompressible

Some Natural Canonical Orderings

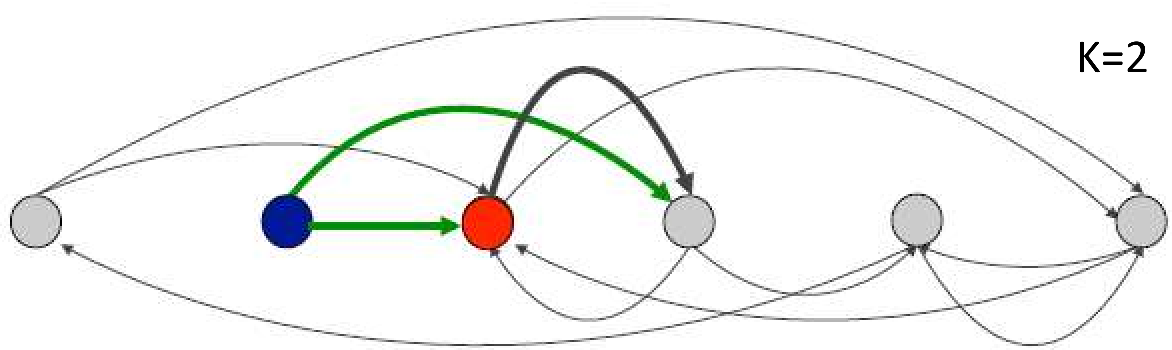
- Random order
- Natural order
  - Time of joining in a social network
  - Lexicographic order of URLs
  - Crawl order
- Graph traversal orders
  - BFS and DFS
- Use attributes of the nodes
  - E.g., Geographic location: order by zip codes
  - May produce a bucket order
- Ties can be broken using more than one order

Shingle Ordered Heuristic

- Obtain a canonical ordering by bringing nodes with similar neighborhoods close together
- Fingerprint neighborhood of each node
- Order the nodes according to the fingerprint
- If fingerprint can capture neighborhood similarity and edge locality, then it will produce good compression via BV, provided the graph is amenable
- Use Jaccard coefficient to measure similarity between nodes
  - $J(A,B) = \frac{|A \cap B|}{|A \cup B|}$
- Double shingle order: break ties within shingle order using a second shingle
- Flickr Graph
  - BV needs 21.8 bits/edge with natural ordering
  - But BV needs only 13.5 bits/edge with shingle ordering

Incompressibility and a Compressible Graph Model

- The following generative models all require $\Omega(\log n)$ bits per edge on average, even if the node labels are removed
  - The preferential attachment model
  - The copying model
  - The evolutionary ACL model [Aiello, Chung, Lu FOCS 2001]
  - Kronecker multiplication model [Leskovec et al PKDD 2005]
  - Model for navigability in social networks [Kleinberg Nature 2000]
- A compressible graph model
  - Begin with a seed graph of nodes with out-degree k, arranged in a cycle
  - Additional nodes arrive in a sequence
  - An arriving node is inserted at a random place in the cycle
    - It links to k-1 out-neighbors of its cycle successor



Locality in the New Model

- If a web designer wants to add a new web page to her website
  - Likely to take some existing web page on her website
  - Modify it as needed (perturbing the set of its outlinks) to obtain the new page
  - Adding a reference to the old web page
  - And publish the new web page on her website
- Since web pages are sorted by URL in our ordering, the old and the new page will be close

Basic Properties of the Model

- Rich gets richer: In the model, in-degrees converge to a power law with exponent $-2-\frac{1}{k-1}$
- High clustering coefficient
- Polynomially many bipartite cliques
- Logarithmic undirected diameter
- Compressible to O(1) bits per edge
  - BV algorithm achieves O(1) bits per edge

Take-away Messages

- Preferential attachment model has power law degree distribution but cannot explain communities well.
- Copying model can explain communities
- Forest fire model leads to densification and diameter shrinkage
- We saw a model to explain formation of network components: GCC, stars, core
- Evolving network model can capture node and edge arrivals with preferential attachment and triangle closing
- Compressible graph model ensures a good compressible graph using BV compression and shingle ordering

Collection of Network Datasets

- http://pajek.imfm.si/doku.php?id=data:urls:index
- http://networkdata.ics.uci.edu/index.html
- http://snap.stanford.edu/data/
- http://www-personal.umich.edu/~mejn/netdata/

Further Reading

- Barabási, A.-L.; R. Albert (1999). "Emergence of scaling in random networks". Science 286 (5439): 509–512
- R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins, and E. Upfal. 2000. Stochastic models for the Web graph. In Proceedings of the 41st Annual Symposium on Foundations of Computer Science (FOCS '00). IEEE Computer Society, Washington, DC, USA
- Jure Leskovec, Deepayan Chakrabarti, Jon M. Kleinberg, Christos Faloutsos: Realistic, Mathematically Tractable Graph Generation and Evolution, Using Kronecker Multiplication. PKDD 2005: 133-145
- Jure Leskovec, Jon M. Kleinberg, Christos Faloutsos: Graphs over time: densification laws, shrinking diameters and possible explanations. KDD 2005: 177-187
- Jure Leskovec, Jon M. Kleinberg, Christos Faloutsos: Graph evolution: Densification and shrinking diameters. TKDD 1(1) (2007)
- Paolo Boldi, Sebastiano Vigna: The webgraph framework I: compression techniques. WWW 2004: 595-602

Preview of Lecture 11: Social Influence Analysis (Part 1)

- Information Diffusion
- Introduction to Social Influence Analysis
- Tests for Social Influence Analysis

Disclaimers

- This course represents opinions of the instructor only. It does not reflect views of Microsoft or any other entity (except of authors from whom the slides have been borrowed).
- Algorithms, techniques, features, etc mentioned here might or might not be in use by Microsoft or any other company.
- Lot of material covered in this course is borrowed from slides across many universities and conference tutorials. These are gratefully acknowledged.

Thanks!