



Big Data on Social Media Mining and Analytics Graph -Network Models

April 3, 2015

About Course Project

Overview

The Course Project is an opportunity for you to apply what you have learned in class to a problem of your interest. There are two project options you can pick from:

Option 1: Your own project

You are encouraged to select a topic and work on your own project. Potential projects usually fall into these two tracks:

- **Applications.** If you're coming to the class with a specific background and interests (e.g. biology, engineering, physics), we'd love to see you apply existing tools to problems related to your particular domain of interest. Pick a real-world problem and apply existing tools to solve it.
- **Models.** You can build a new model (algorithm), or a new variant of existing models, and apply it to tackle social media tasks. This track might be more challenging, and sometimes leads to a piece of publishable work.

Don't be afraid to think outside of the box.

About Course Project

Option 2: 国科大-社交网络分析平台和系统

- S 结构分析
- C 内容分析
- T 传播分析
- E 事件检测
- V 可视化分析
- H 高效索引查询
- I 影响力分析
- P 用户行为分析

系统开发单位	系统名称 ^{a)}	项目来源	平台/系统功能 ^{b)}							
			S	C	T	E	V	H	I	P
斯坦福大学	SNAP	美国自然科学基金、微软、雅虎	●	○	○	○	●	○	○	○
卡内基梅隆大学	Pegasus	美国自然科学基金、雅虎、劳伦斯·利弗莫尔实验室	●	○	●	○	●	●	●	○
卡内基梅隆大学	AutoMap	美国自然科学基金、美国陆军研究院	●	●	●	●	○	○	●	●
卡内基梅隆大学	OddBall	美国国家科学基金、美国能源部的国家实验室	●	○	●	●	●	○	○	○
西北大学(美国)	C-INKNOW	美国自然科学基金、美国国家健康以及陆军研究院	●	●	○	○	●	○	○	●
华盛顿大学	Statnet	美国国家科学基金、海军研究局	●	○	○	●	●	○	○	●
印第安纳大学	NWB	美国自然科学基金	●	○	●	○	●	○	○	○
匹兹堡大学	EpiFast	美国国家科学基金网络技术与系统	●	○	●	○	●	●	○	○
加州大学洛杉矶分校	HCS		●	○	●	○	●	●	●	●
纽约州立大学石溪分校	NetworkX		●	○	●	○	●	○	○	○
罗兰大学	CFinder	匈牙利自然科学基金	●	○	●	○	●	○	●	●
Google	Pregel		●	○	●	○	○	●	○	○
IBM	X-RIME		●	○	●	○	●	●	○	○
微软研究院	CoSBI Lab Graph		●	○	●	○	●	●	●	○
Gephi.org	Gephi		●	○	●	○	●	○	●	○
Analytic Technologies	UCINET		●	○	●	○	●	○	●	○
ATOS	MyInfo+	伦敦奥运会合作项目	●	●	●	●	●	●	○	○
开源项目	Pajek		●	○	●	○	●	○	○	○
开源项目	GraphLab		●	●	●	○	○	○	○	○
清华大学	SAE	国家高技术研究发展计划、国家自然科学基金及华为支持项目	●	●	●	○	○	●	●	●

Some previous cases

Title

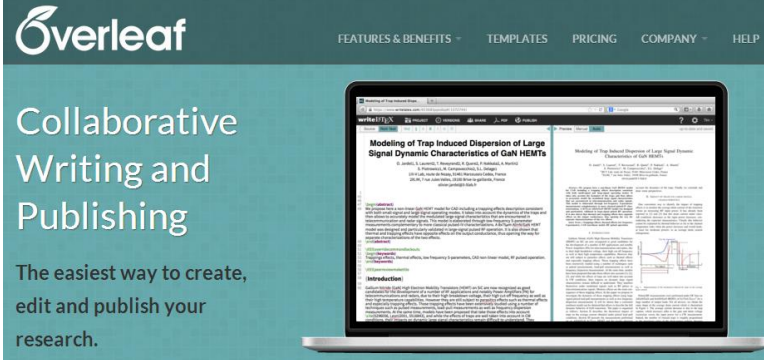
Name, Affiliation

Abstract

1. Introduction (Motivation)
2. Related Work
3. Methodology
4. Experiment (Datasets, Results and Analysis)
5. Conclusion and Future work (Contribution)
6. References

Top Conferences

KDD ICDM WSDM
IJCAI AAAI
WWW CIKM

The image shows the Overleaf website banner. It features the Overleaf logo in the top left corner. To the right of the logo are navigation links: FEATURES & BENEFITS, TEMPLATES, PRICING, COMPANY, and HELP. The main text reads "Collaborative Writing and Publishing" in a large, bold font. Below this, a smaller line of text says "The easiest way to create, edit and publish your research." To the right of the text is a laptop displaying a document with a title "Modeling of Trap Induced Dispersion of Large Signal Dynamic Characteristics of GaN HEMTs" and a plot of a signal waveform.

Piazza for Q&A

The screenshot displays the Piazza Q&A interface for CS 224W. The top navigation bar includes links for 'Q & A', 'Resources', and 'Statistics', along with a user profile for Henry Wang. Below the navigation bar, a horizontal list of topics is shown, including 'hw1', 'hw2', 'hw3', 'hw4', 'project', 'logistics', 'other', 'hw0', 'snap.py', 'snap', and 'python'. The main content area is divided into three sections: a sidebar on the left, a central question area, and a follow-up discussion section at the bottom.

Sidebar: A list of questions with their respective counts and status icons. The questions include:

- Suggestion for poster printing (3)
- Export part of a graph in GEPHI (5)
- Project report submission blue box in ... (1)
- Equal Contributions of Group Members (3)
- Code submission vs pdf submission (9)
- Level of detail in mathematical analysis (4)
- Print posters (3)
- Poster board question (4)
- Collaborative LaTeX (3)
- Instr Hima's weekend OH 6th dec... (6)
- Office hours today (6 Dec) (3)
- WEEK 11/30 - 12/6
- Goodness of Fit (2)
- Instr Hima's OH at Gates 448 (3p... (3)
- Number of Simulations (3)
- Instr Vikesh's OH in Gates B24A 2... (3)
- Language Distance Graph from 12/4's I... (2)

Central Question Area: The selected question is 'Project grades' with 116 views. The question text is 'When can we expect the final project grades?' and 'Thanks.' The question was updated 3 months ago by Arun Mahendra. Below the question is a section for 'the students' answer', where a student has responded: 'I can see the project grades (letter) on Axxess.' This answer was updated 3 months ago by Anonymous.

Follow-up Discussion: A section for 'followup discussions' for lingering questions and comments. It includes a filter for 'Resolved' and 'Unresolved' questions. A discussion by Sajid Zaidi, 3 months ago, asks 'Can we see the cutoffs for overall class grades?'.

Why should I use network models?

- In may 2011, Facebook had 721 millions users. A Facebook user at the time had an average of 190 users -> a total of 68.5 billion friendships
 - What are the principal underlying processes that help initiate these friendships
 - How can these seemingly independent friendships form this complex friendship network?
- In social media there are many networks with millions of nodes and billions of edges.
 - They are complex and it is difficult to analyze them

So, what do we do?

- We design models that generate, on a smaller scale, graphs similar to real-world networks.
- Hoping that these models simulate properties observed in real-world networks well, the analysis of real-world networks boils down to a cost-efficient measuring of different properties of simulated networks
 - Allow for a better understanding of phenomena observed in real-world networks by providing concrete mathematical explanations; and
 - Allow for controlled experiments on synthetic networks when real-world networks are not available.
- These models are designed to accurately model properties observed in real-world networks

Properties of Real-World Networks

Power-law Distribution, High Clustering Coefficient, Small Average Path Length

Degree Distribution

Degree Distribution

- Consider the distribution of wealth among individuals. Most individuals have average capitals, whereas a few are considered wealthy. In fact, we observe exponentially more individuals with average capital than the wealthier ones.
- Similarly, consider the population of cities. Often, a few metropolitan areas are densely populated, whereas other cities have an average population size.
- In social media, we observe the same phenomenon regularly when measuring popularity or interestingness for entities.

Degree Distribution

- Many sites are visited less than a 1,000 times a month whereas a few are visited more than a million times daily.
- Social media users are often active on a few sites whereas some individuals are active on hundreds of sites.
- There are exponentially more modestly priced products for sale compared to expensive ones.
- There exist many individuals with a few friends and a handful of users with thousands of friends

(Degree Distribution)

Power Law Distribution

- When the frequency of an event changes as a power of an attribute -> the frequency follows a **power-law**
- Let $p(k)$ denote the fraction of individuals having degree k .

$$p_k = ak^{-b},$$

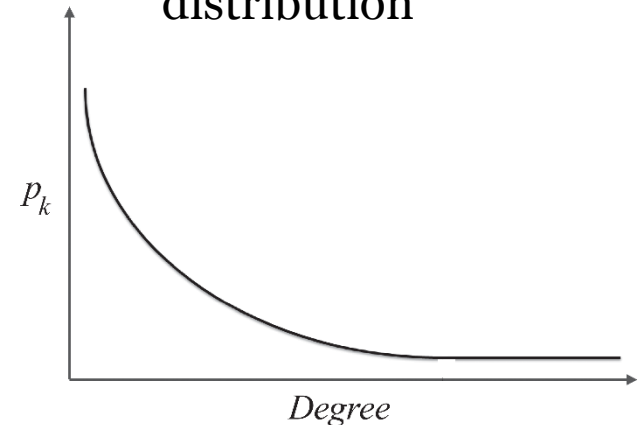
$$\ln p_k = -b \ln k + \ln a.$$

b: the power-law exponent and its value is typically in the range of $[2, 3]$
a: power-law intercept

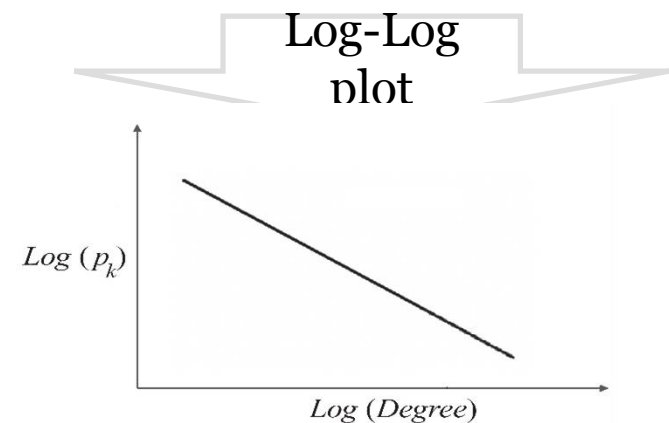
Power Law Distribution

- Many real-world networks exhibit a *power-law* distribution.
- Power laws seem to dominate in cases where the quantity being measured can be viewed as a type of **popularity**.
- A power-law distribution implies that small occurrences are common, whereas large instances are extremely rare

A typical shape of a power-law distribution



(a) Power-Law Degree Distribution



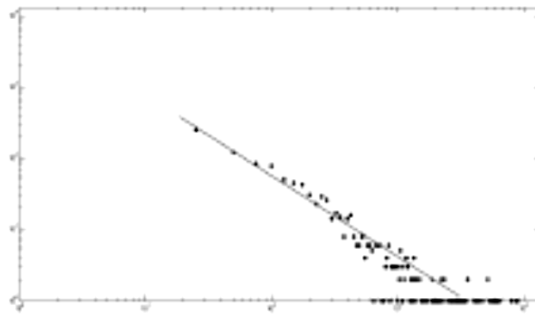
(b) Log-Log Plot of Power-Law Degree Distribution

Power-law Distribution: Test

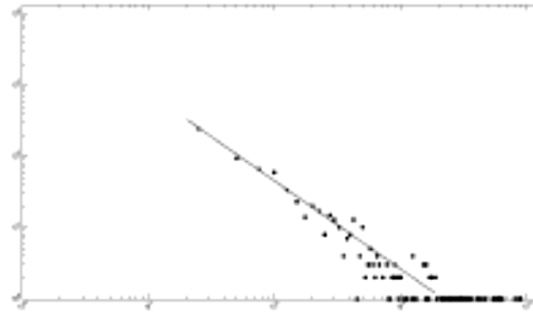
- Test whether a network exhibits a power-law distribution
- Pick a popularity measure and compute it for the whole network. For instance, we can take the number of friends in a social network
 - Compute $p(k)$, the fraction of individuals having popularity k .
 - Plot a log-log graph, where the x-axis represents $\ln k$ and the y-axis represents $\ln p(k)$.
 - If a power-law distribution exists, we should observe a straight line
- The results can be inaccurate

Power-Law Distribution: Real-World Networks

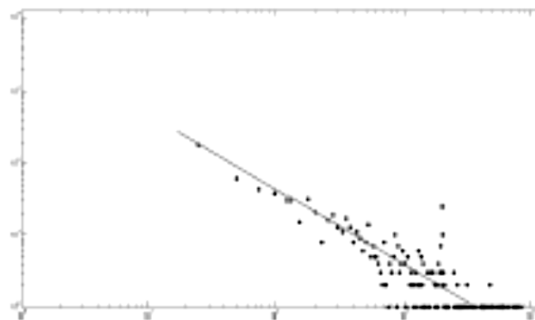
- Networks with power-law degree distribution are often called **scale-free** networks



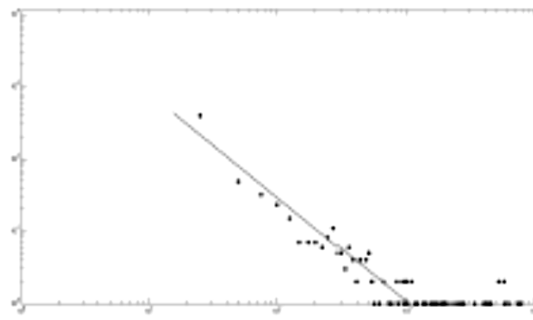
(a) BlogCatalog



(b) MyBlogLog



(c) Twitter



(d) MySpace

Clustering Coefficient

Clustering Coefficient

- In real-world networks, friendships are highly transitive, i.e., friends of an individual are often friends with one another
 - These friendships form triads -> high average [local] clustering coefficient
- In May 2011, Facebook had an average clustering coefficient of 0.5 for individuals who had 2 friends.

Web	Facebook	Flickr	LiveJournal	Orkut	YouTube
0.081	0.14 (with 100 friends)	0.31	0.33	0.17	0.13

Average Path Length

The Average Shortest Path

- In real-world networks, any two members of the network are usually connected via short paths. In other words, the average path length is small
 - Six degrees of separation:
 - **Stanley Milgram** In the well-known small-world experiment conducted in the 1960's conjectured that people around the world are connected to one another via a path of at most 6 individuals
 - Four degrees of separation:
 - **Lars Backstrom et al.** in May 2011, the average path length between individuals in the Facebook graph was 4.7. (4.3 for individuals in the US)

Web	Facebook	Flickr	LiveJournal	Orkut	YouTube
16.12	4.7	5.67	5.88	4.25	5.10

Stanley Milgram's Experiments

- Random people from Nebraska were asked to send a letter (via intermediaries) to a stock broker in Boston
- S/he could only send to someone with whom they were on a first-name basis

Among the letters that reached the target, the average path length was six.



Stanley Milgram (1933-1984)

Random Graphs

Random Graphs

- We start with the most basic assumption on how friendships are formed.

Random Graph's main assumption:

Edges (i.e., friendships) between nodes (i.e., individuals) are formed randomly.

Random Graph Model – $G(n,p)$

- We discuss two random graph models
- Formally, we can assume that for a graph with a fixed number of nodes n , any of the $\binom{n}{2}$ edges can be formed independently, with probability p . This graph is called a random graph and we denote it as $G(n, p)$ model.
 - This model was first proposed independently by Edgar Gilbert and Solomonoff and Rapoport.

$C(n, 2)$ or $\binom{n}{2}$ is # of combinations of two objects from a set of n objects

Random Graph Model - $G(n,m)$

- Another way of randomly generating graphs is to assume both number of nodes n and number of edges m are fixed. However, we need to determine which m edges are selected from the set of $\binom{n}{2}$ possible edges
 - Let Ω denote the set of graphs with n nodes and m edges
 - There are $|\Omega|$ different graphs with n nodes and m edges

$$|\Omega| = \binom{\binom{n}{2}}{m}$$

- To generate a random graph, we uniformly select one of the $|\Omega|$ graphs (the selection probability is $1/|\Omega|$)

This model proposed first by Paul Erdos and Alfred Renyi

Modeling Random Graphs, Cont.

- In the limit (when n is large), both models ($\mathbf{G(n, p)}$ and $\mathbf{G(n, m)}$) act similarly
 - The expected number of edges in $G(n, p)$ is $\binom{n}{2}p$
 - We can set $\binom{n}{2}p = m$ and in the limit, we should get similar results.

Differences:

- The $G(n, m)$ model contains a fixed number of edges
- The $G(n, p)$ model is likely to contain none or all possible edges

Expected Degree

The expected number of edges connected to a node (expected degree) in $G(n, p)$ is $c=(n - 1)p$

- **Proof:**

- A node can be connected to at most $n-1$ nodes (or $n-1$ edges)
- All edges are selected independently with probability p
- Therefore, on average, $(n - 1)p$ edges are selected

- $C=(n-1)p$ or equivalently,

$$p = \frac{c}{n - 1}.$$

Expected Number of Edges

- The expected number of edges in $G(n, p)$ is $\binom{n}{2}p$
- **Proof:**
 - Since edges are selected independently, and we have a maximum $\binom{n}{2}$ edges, the expected number of edges is $p\binom{n}{2}$

The probability of Observing m edges

Given the $G(n, p)$ process, the probability of observing m edges is binomial distribution

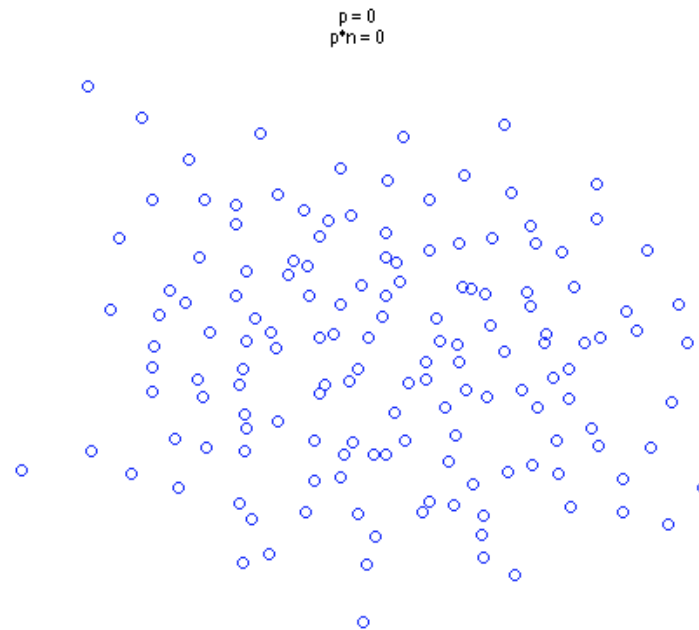
$$P(|E| = m) = \binom{\binom{n}{2}}{m} p^m (1 - p)^{\binom{n}{2} - m}$$

- **Proof:**

- m edges are selected from the $\binom{n}{2}$ possible edges.
- These m edges are formed with probability p^m and other edges are not formed (to guarantee the existence of only m edges) with probability

$$(1 - p)^{\binom{n}{2} - m}$$

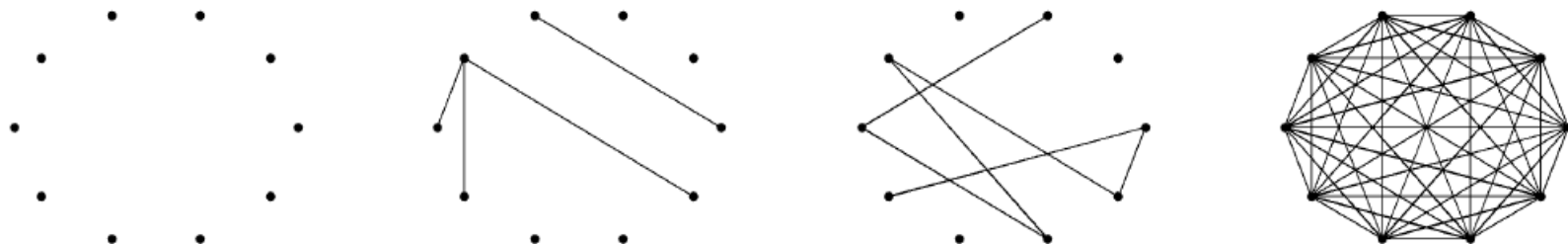
Evolution of Random Graphs



The Giant Component

- In random graphs, when nodes form connections, after some time, a large fraction of nodes get connected, i.e., there is a path between any pair of them.
- This large fraction forms a connected component, commonly called the **largest connected component** or the **giant component**.
- In random graphs:
 - $p = 0 \rightarrow$ the size of the giant component is 0
 - $p = 1 \rightarrow$ the size of the giant component is n

The Giant Component



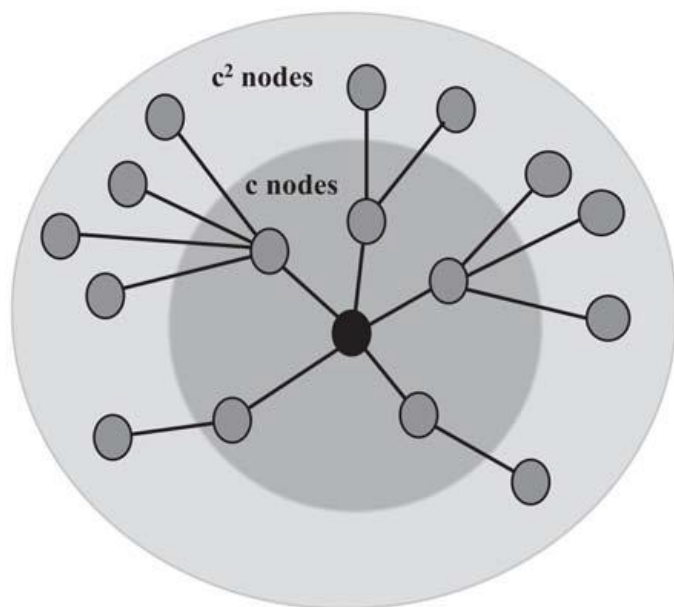
Probability (p)	0.0	0.055	0.11	1.0
Average node degree (c)	0.0	0.8	~ 1	$n-1 = 9$
Diameter	0	2	6	1
Giant component size	0	4	7	10
Average path length	0.0	1.5	2.66	1.0

Evolution of Random Graphs. Here, p is the random graph generation probability, c is the average degree, ds is the diameter size, slc is the size of the largest component, and l is the average path length. The **highlighted column** denotes phase transition in the random graph.

Phase Transition

- The point where diameter value starts to shrink in a random graph is called ***Phase Transition***.
- In a random graph, *phase transition* happens when average node degree, $c = 1$, or when $p = 1/(n-1)$
- At the point of Phase Transition, the following phenomena are observed:
 - The giant component that just started to appear, starts grow, and
 - The diameter that just reached its maximum value, starts decreasing.

Why $C=1$?



Nodes Visited by Moving n -hops away in a Random Graph.
 c denotes the expected node degree.

Proof. (Sketch) Consider a random graph with expected node degree c , where $c = p(n-1)$. In this graph, consider any **connected** set of nodes S and consider the complement set $\bar{S} = V - S$. For the sake of our proof, we assume that $|S| \ll |\bar{S}|$. Given any node v in S , if we move one hop (edge) away from v , we visit approximately c nodes. Following the same argument, if we move one hop away from nodes in S , we visit approximately $|S|c$ nodes. Assuming $|S|$ is small, the nodes in S only visit nodes in \bar{S} , and when moving one hop away from S , the set of nodes “guaranteed to be connected” gets larger by a factor c . The connected set of visited nodes gets c^2 times larger when moving two hops and so on. Now, in the limit, if we want this component of visited nodes to become the largest connected component, then after traveling n hops, we must have

$$c^n \geq 1 \text{ or equivalently } c \geq 1.$$

Otherwise (i.e., $c < 1$), the number of visited nodes dies out exponentially. Hence, phase transition happens at $c = 1$.² \square

Properties of Random Graphs

Degree Distribution

- When computing degree distribution, we estimate the probability of observing $P(d_v = d)$ for node v
- For a random graph generated by $G(n, p)$ this probability is

$$P(d_v = d) = \binom{n-1}{d} p^d (1-p)^{n-1-d},$$

- This is a binomial degree distribution. In the limit this will become the Poisson degree distribution

Expected Local Clustering Coefficient

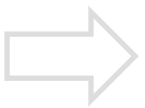
The expected local clustering coefficient for node v of a random graph generated by $G(n, p)$ is p

- **Proof:**

$$C(v) = \frac{\text{number of } v\text{-neighbor pairs that are connected}}{\text{number of } v\text{-neighbor pairs}}$$

- v can have different degrees depending on the random procedure so the expected value is,

$$E(C(v)) = \sum_{d=0}^{d=n-1} E(C(v)|d_v = d)P(d_v = d)$$



Expected Local Clustering Coefficient, Cont.

$$\mathbf{E}(C(v)) = \sum_{d=0}^{d=n-1} \mathbf{E}(C(v)|d_v = d)P(d_v = d)$$



$$\mathbf{E}(C(v)|d_v = d) = \frac{\text{number of } v \text{ neighbor pairs that are connected}}{\text{number of } v \text{ neighbor pairs}} = \frac{p \binom{d}{2}}{\binom{d}{2}} = p$$



$$\mathbf{E}(C(v)) = p \sum_{d=0}^{d=n-1} P(d_v = d) = p$$

Global Clustering Coefficient

The global clustering coefficient of a random graph generated by $G(n, p)$ is p

- **Proof:**

- The global clustering coefficient of any graph defines the probability of two neighbors of the same node that are connected.
- In a random graph, for any two nodes, this probability is the same and is equal to the generation probability p that determines the probability of two nodes getting connected

The Average Path Length

The average path length in a random graph is $l \approx \frac{\ln |V|}{\ln c}$.

Proof. (Sketch) The proof is similar to the proof provided in determining when phase transition happens.

Let \mathcal{D} denote the expected diameter size in the random graph. Starting with any node in a random graph and its expected degree c , one can visit approximately c nodes by traveling one edge, c^2 nodes by traveling two edges, and $c^{\mathcal{D}}$ nodes by traveling “diameter” number of edges. After this step, almost all nodes should be visited. In this case, we have

$$c^{\mathcal{D}} \approx |V|.$$

In random graphs, the expected diameter size tends to the average path length l in the limit.

$$c^{\mathcal{D}} \approx c^l \approx |V|.$$

Taking the logarithm from both sides we get $l \approx \frac{\ln |V|}{\ln c}$. Therefore, the average path length in a random graph is equal to $\frac{\ln |V|}{\ln c}$. \square

Modeling Real-World Networks with Random Graphs

- Compute the average degree c in the given network, then compute p , by using: $c/(n-1) = p$, then generate the random graph.
- How good is the model?
 - random graphs perform well in modeling the average path lengths; however, when considering the *transitivity*, the random graph model drastically underestimates the clustering coefficient.

	Original Network				Simulated Random Graph	
Network	Size	Average Degree	Average Path Length	C	Average Path Length	C
Film Actors	225,226	61	3.65	0.79	2.99	0.00027
Medline Coauthorship	1,520,251	18.1	4.6	0.56	4.91	1.8×10^{-4}
E.Coli	282	7.35	2.9	0.32	3.04	0.026
C.Elegans	282	14	2.65	0.28	2.25	0.05

Small-World Model

The assumption behind the random graph model is that connections in real-world networks are formed at random. Although unrealistic, random graphs can model average path lengths in real-world networks properly, but underestimate the clustering coefficient. To mitigate this problem, Duncan J. Watts and Steven Strogatz in 1997 proposed the small-world model.

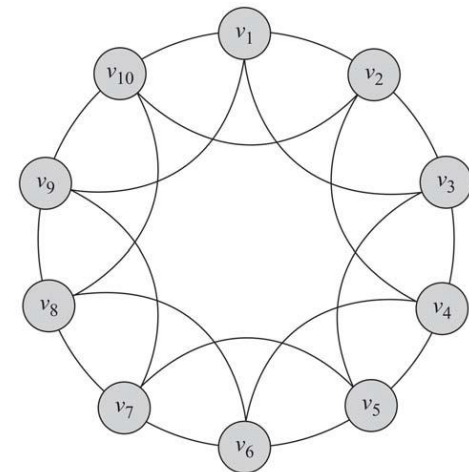
Small-world Model

- Small-world Model also known as **the Watts and Strogatz model** is a special type of random graphs with small-world properties, including:
 - Short average path length and;
 - High clustering.
- It was proposed by Duncan J. Watts and Steven Strogatz in their joint 1998 Nature paper

Small-world Model

- In real-world interactions, many individuals have a limited and often at least, a fixed number of connections
- In graph theory terms, this assumption is equivalent to embedding individuals in a regular network.
- A regular (ring) lattice is a special case of regular networks where there exists a certain pattern on how ordered nodes are connected to one another.
- In particular, in a regular lattice of degree c , nodes are connected to their previous $c/2$ and following $c/2$ neighbors. Formally, for node set $V = \{v_1, v_2, v_3, \dots, v_n\}$, an edge exists between node i and j if and only if

$$0 < |i - j| \leq c/2.$$



Regular Lattice
of Degree 4

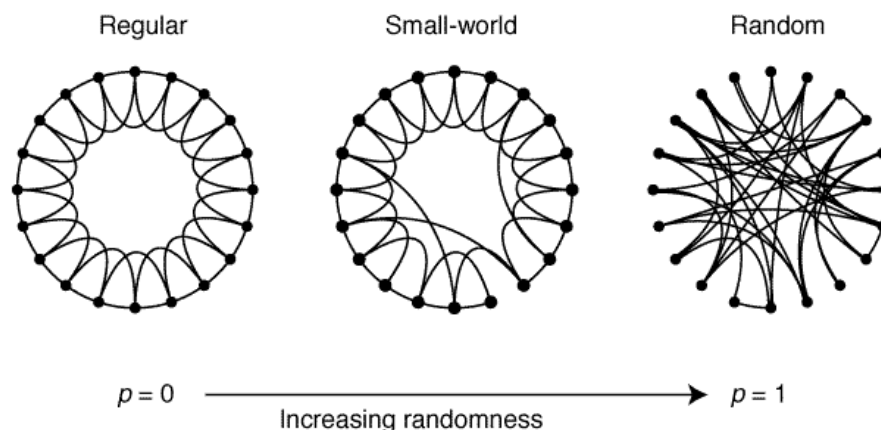
The regular lattice can model transitivity well; however, the average path length is too high.

Constructing Small World Networks

Algorithm 4.1 Small-World Generation Algorithm

Require: Number of nodes $|V|$, mean degree c , parameter β

- 1: **return** A small-world graph $G(V, E)$
 - 2: $G =$ A regular ring lattice with $|V|$ nodes and degree c
 - 3: **for** node v_i (starting from v_1), and all edges $e(v_i, v_j), i < j$ **do**
 - 4: $v_k =$ Select a node from V uniformly at random.
 - 5: **if** rewiring $e(v_i, v_j)$ to $e(v_i, v_k)$ does not create loops in the graph or multiple edges between v_i and v_k **then**
 - 6: rewire $e(v_i, v_j)$ with probability β : $E = E - \{e(v_i, v_j)\}, E = E \cup \{e(v_i, v_k)\}$;
 - 7: **end if**
 - 8: **end for**
 - 9: **Return** $G(V, E)$
-



Small-World Model Properties

Degree Distribution

- The degree distribution for the small-world model is

$$P(d_v = d) = \sum_{n=0}^{\min(d-c/2, c/2)} \binom{c/2}{n} (1-\beta)^n \beta^{c/2-n} \frac{(\beta c/2)^{d-c/2-n}}{(d-c/2-n)} e^{-\beta c/2},$$

- In practice, in the graph generated by the small world model, most nodes have similar degrees due to the underlying lattice.

Regular Lattice and Random Graph: Clustering Coefficient and Average Path Length

- Regular Lattice:
 - Clustering Coefficient (high):

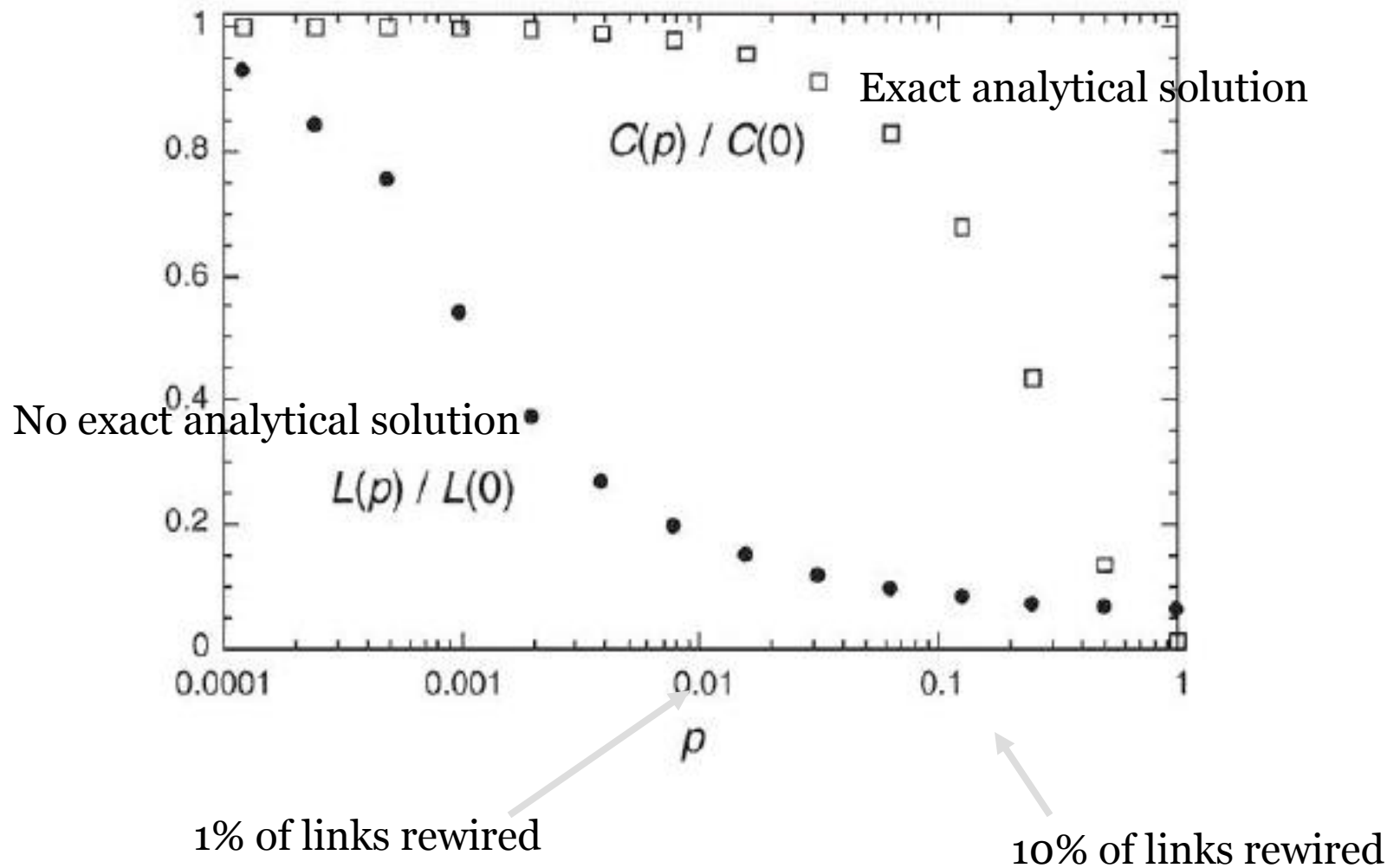
$$\frac{3(c-2)}{4(c-1)} \approx \frac{3}{4}$$

- Average Path Length (high): $n/2c$
- Random Graph:
 - Clustering Coefficient (low): p
 - Average Path Length (ok!) : $\ln |V| / \ln c$

What happens in Between?

- Does smaller average path length mean smaller clustering coefficient?
- Does larger average path length mean larger clustering coefficient?
- Through numerical simulation
 - As we increase p from 0 to 1
 - Fast decrease of average distance
 - Slow decrease in clustering coefficient

Change in Clustering Coefficient and Average Path Length as a Function of the Proportion of Rewired Edges



Clustering Coefficient for Small-world model with rewiring

- The probability that a connected triple stays connected after rewiring consists of two parts
 1. The probability that none of the 3 edges were rewired is $(1-p)^3$
 2. The probability that other edges were rewired back to form a connected triple is very small and can be ignored
- Clustering coefficient

$$C(p) \approx (1 - p)^3 C(0).$$

p

Modeling Real-World Networks with the Small-World Model

- Given a real-world network in which average
- degree c and clustering coefficient C is given, we set $C(p) = C$ and determine β ($=p$) using equation

$$C(p) \approx (1 - p)^3 C(0).$$

- Given β , c , and n (size of the real-world network), we can simulate the small-world model.

Real-World Network and Simulated Graphs

	Original Network				Simulated Graph	
Network	<i>Size</i>	<i>Average Degree</i>	<i>Average Path Length</i>	<i>C</i>	<i>Average Path Length</i>	<i>C</i>
Film Actors	225,226	61	3.65	0.79	4.2	0.73
Medline Coauthorship	1,520,251	18.1	4.6	0.56	5.1	0.52
E.Coli	282	7.35	2.9	0.32	4.46	0.31
C.Elegans	282	14	2.65	0.28	3.49	0.37

Preferential Attachment Model

Barabasi-Albert (BA) model

When new nodes are added to networks, they are more likely to connect to existing nodes that many others have connected to.

Preferential Attachment: An Example

- Networks:
 - When a new user joins the network, the probability of connecting to existing nodes is proportional to the nodes' degree
- Distribution of wealth in the society:
 - The rich get richer
 - Unlike random graphs in which we assume friendships are formed randomly, in the preferential attachment model we assume that individuals are more likely to befriend gregarious others.

Constructing Scale-free Networks

- Graph $G(V_0, E)$ is given
- For any new node v to the graph
 - Connect v to a random node $v_i \in V_0$, with probability $P(v_i) = \frac{d_i}{\sum_j d_j}$.

Preferential Attachment

Require: Graph $G(V_0, E_0)$, where $|V_0| = m_0$ and $d_v \geq 1 \forall v \in V_0$, number of expected connections $m \leq m_0$, time to run the algorithm t

```
1: return A scale-free network
2: //Initial graph with  $m_0$  nodes with degrees at least 1
3:  $G(V, E) = G(V_0, E_0)$ ;
4: for 1 to  $t$  do
5:    $V = V \cup \{v_i\}$ ; // add new node  $v_i$     one at a time
6:   while  $d_i \neq m$  do
7:     Connect  $v_i$  to a random node  $v_j \in V, i \neq j$  ( i.e.,  $E = E \cup \{e(v_i, v_j)\}$  )
       with probability  $P(v_j) = \frac{d_j}{\sum_k d_k}$ .
8:   end while
9: end for
10: Return  $G(V, E)$ 
```

Intrinsically, higher degree nodes get more attention from newly added nodes.

Preferential Attachment Model

The model incorporates two ingredients

- (1) the *growth* element and
- (2) the *preferential attachment* element – to achieve a scale-free network.

The *growth* is realized by adding nodes as time goes by. The *preferential attachment* is realized by connecting to node v_i based on its degree probability.

Generating networks with

- A power-law degree distribution.
- Small average path length.
- fail to high clustering coefficients

Properties of the Preferential Attachment Model

Properties

- Degree Distribution:

$$P(d) = \frac{2m^2}{d^3},$$

- Clustering Coefficient:

$$C = \frac{m_0 - 1}{8} \frac{(\ln t)^2}{t},$$

- Average Path Length:

$$l \sim \frac{\ln |V|}{\ln(\ln |V|)}.$$

Similar to random graphs, we can simulate real-world networks by generating a preferential attachment model by setting the expected degree m

A Comparison between Real-World Networks and Simulated Graphs using Preferential Attachment. C denotes the average clustering coefficient. The last two columns show the average path length and the clustering coefficient for the preferential-attachment graph simulated for the real-world network. Note that average path lengths are modeled properly, whereas the clustering coefficient is underestimated

	Original Network				Simulated Graph	
Network	Size	Average Degree	Average Path Length	C	Average Path Length	C
Film Actors	225,226	61	3.65	0.79	4.90	≈ 0.005
Medline Coauthorship	1,520,251	18.1	4.6	0.56	5.36	≈ 0.0002
E.Coli	282	7.35	2.9	0.32	2.37	0.03
C.Elegans	282	14	2.65	0.28	1.99	0.05