Networks and Random Graphs

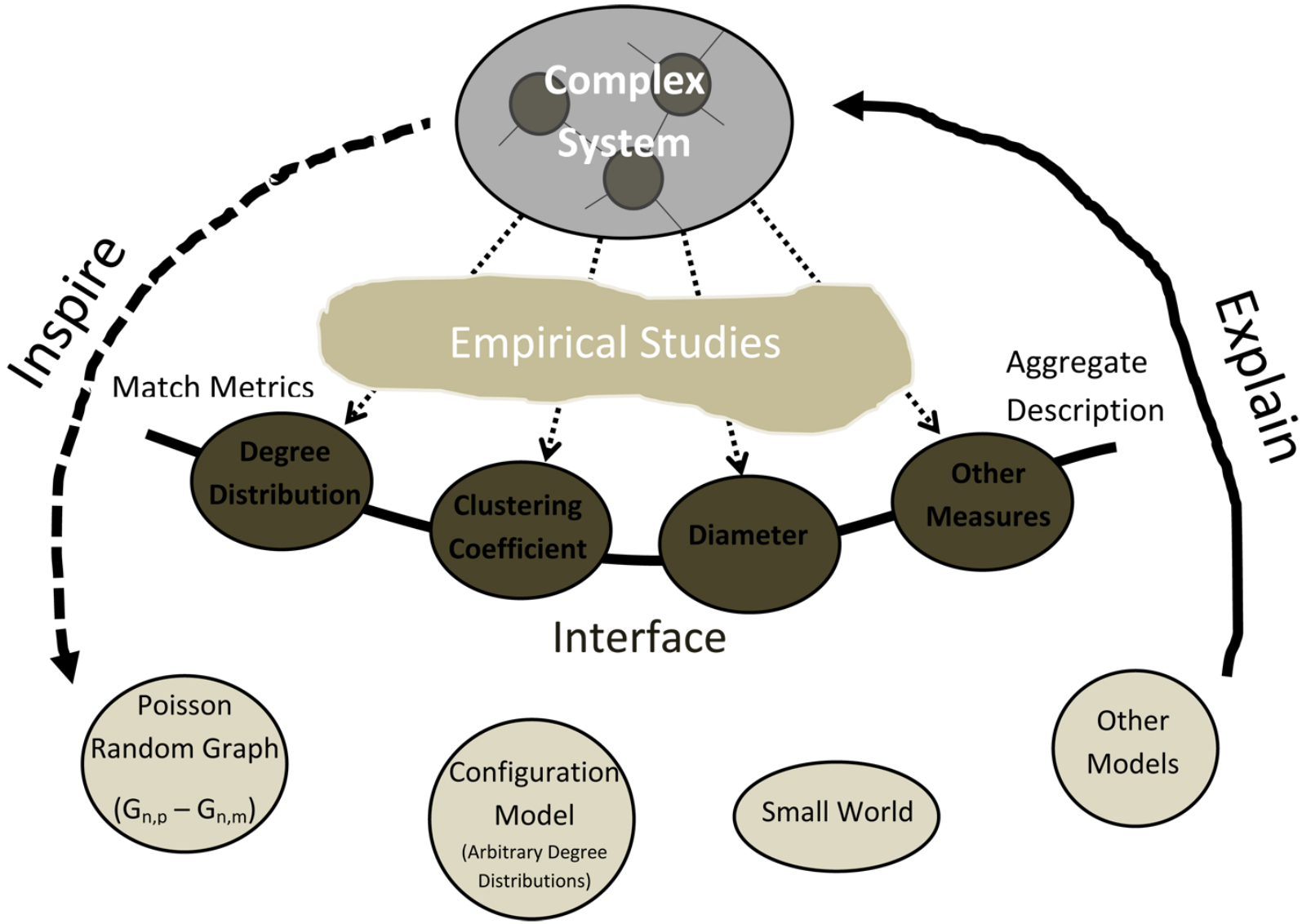
Modelling Real World Networks

Paris Siminelakis

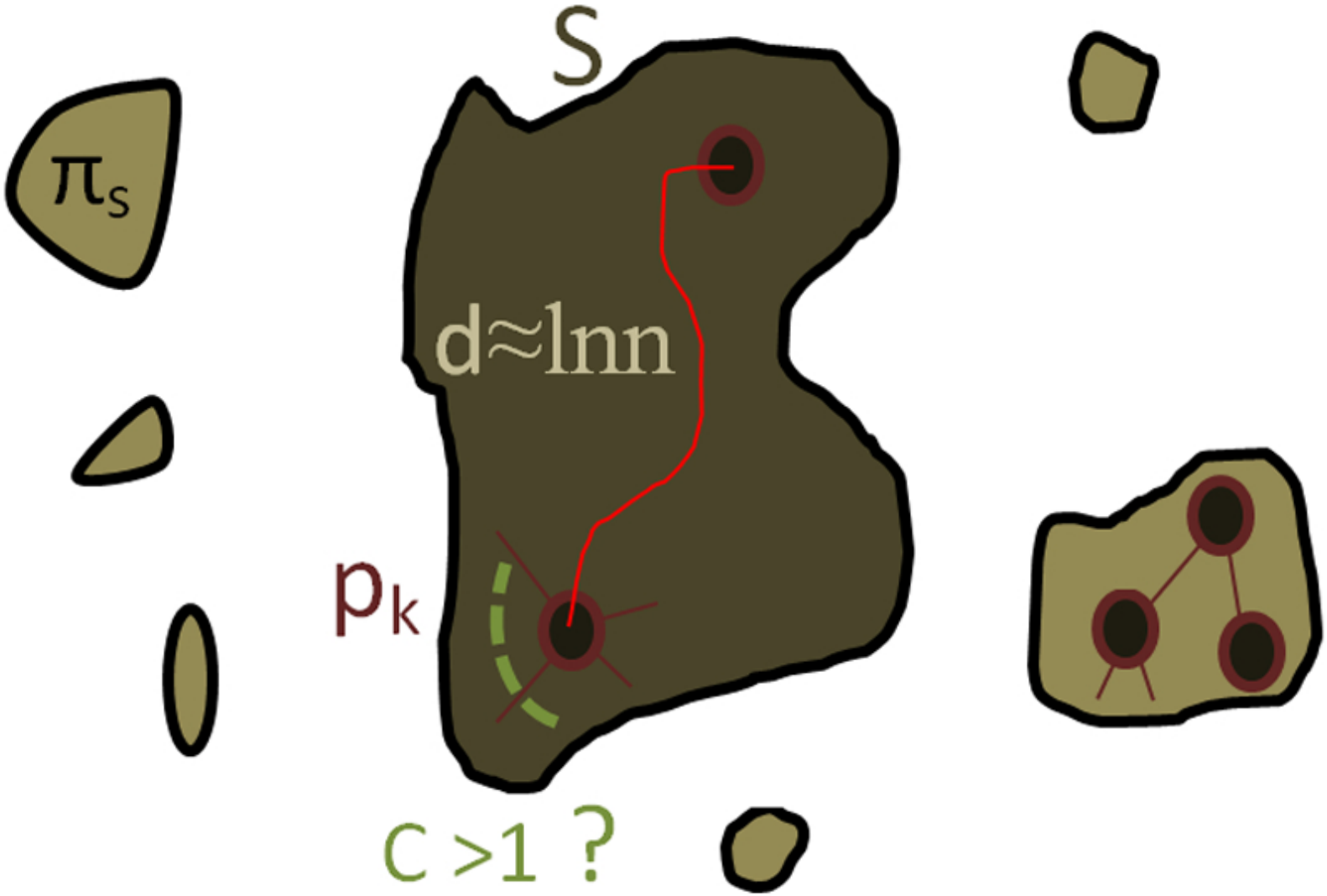
School of Electrical and Computer Engineering at the National Technical University of Athens

21/11/2011

Networks Science Overview



The Poisson Random Graph



The Next Step

We have seen that $G_{n,p}$ isn't a realistic model for real world networks:

- x Poisson Degree Distribution $\neq$ Power Law
- x Low Transitivity and Clustering
- x Degrees uncorrelated

Need for new models, that closely match real world network properties:

- **Configuration Model**
- Preferential Attachment
- Exponential Random Graphs

The Next Step

Main themes utilized by scientists to model networks:

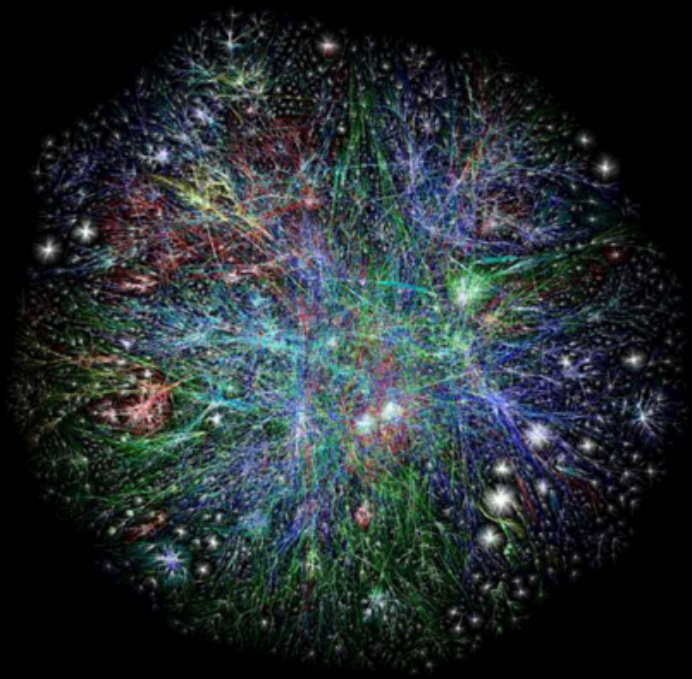
- Randomness
- Feedback Mechanisms
- **Simple rules**

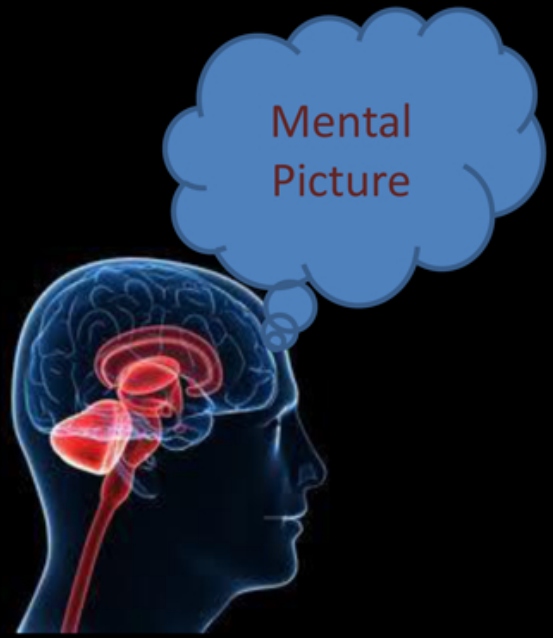
We aim to provide models that:

- Capture Qualitative Characteristics
- Provide Natural Generative and Evolution Mechanisms.
- Analytically **Tractable**.

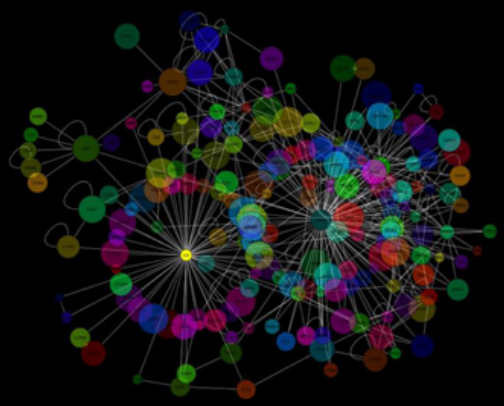
Network Metrics

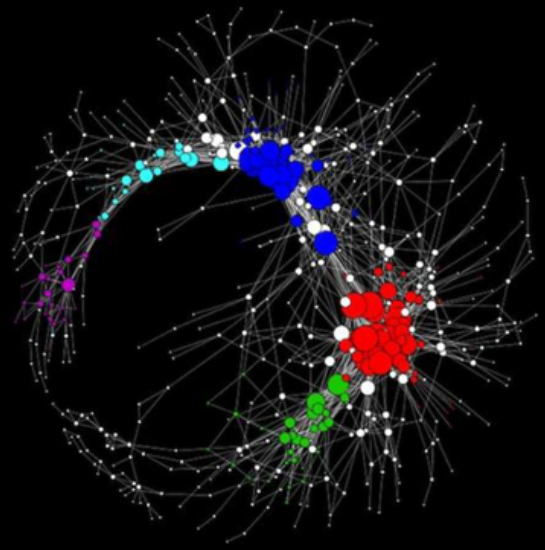
Concise Qualitative Description

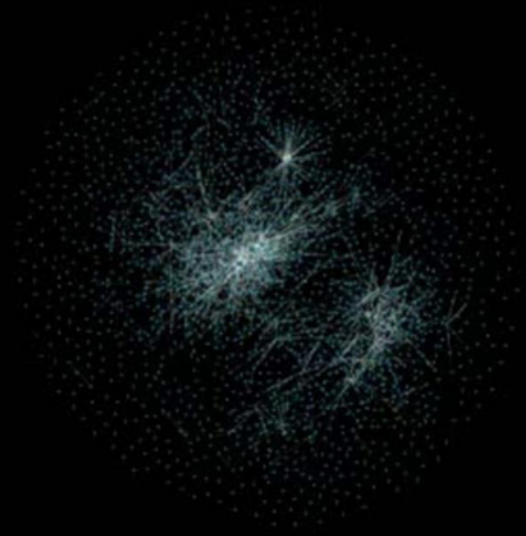


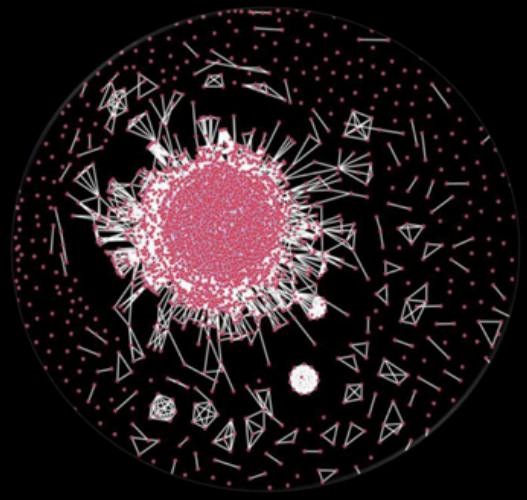


Similarities and Differences



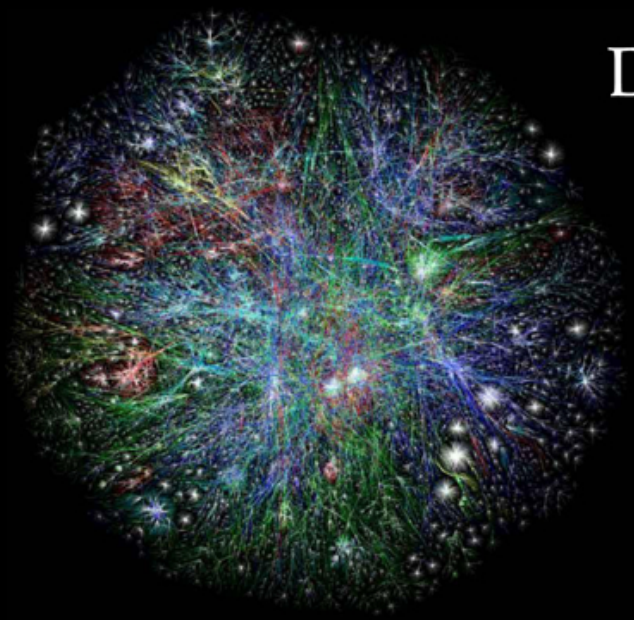






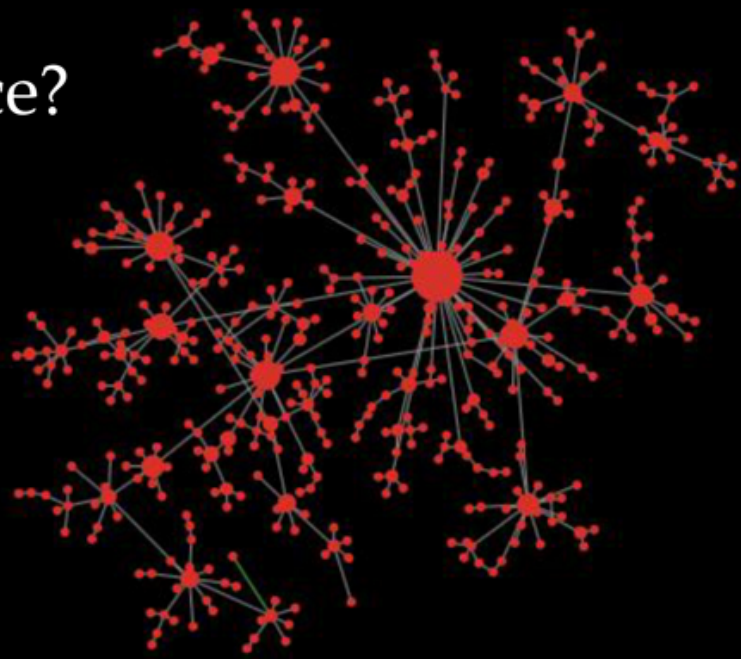
Evaluate our Models

Internet Graph

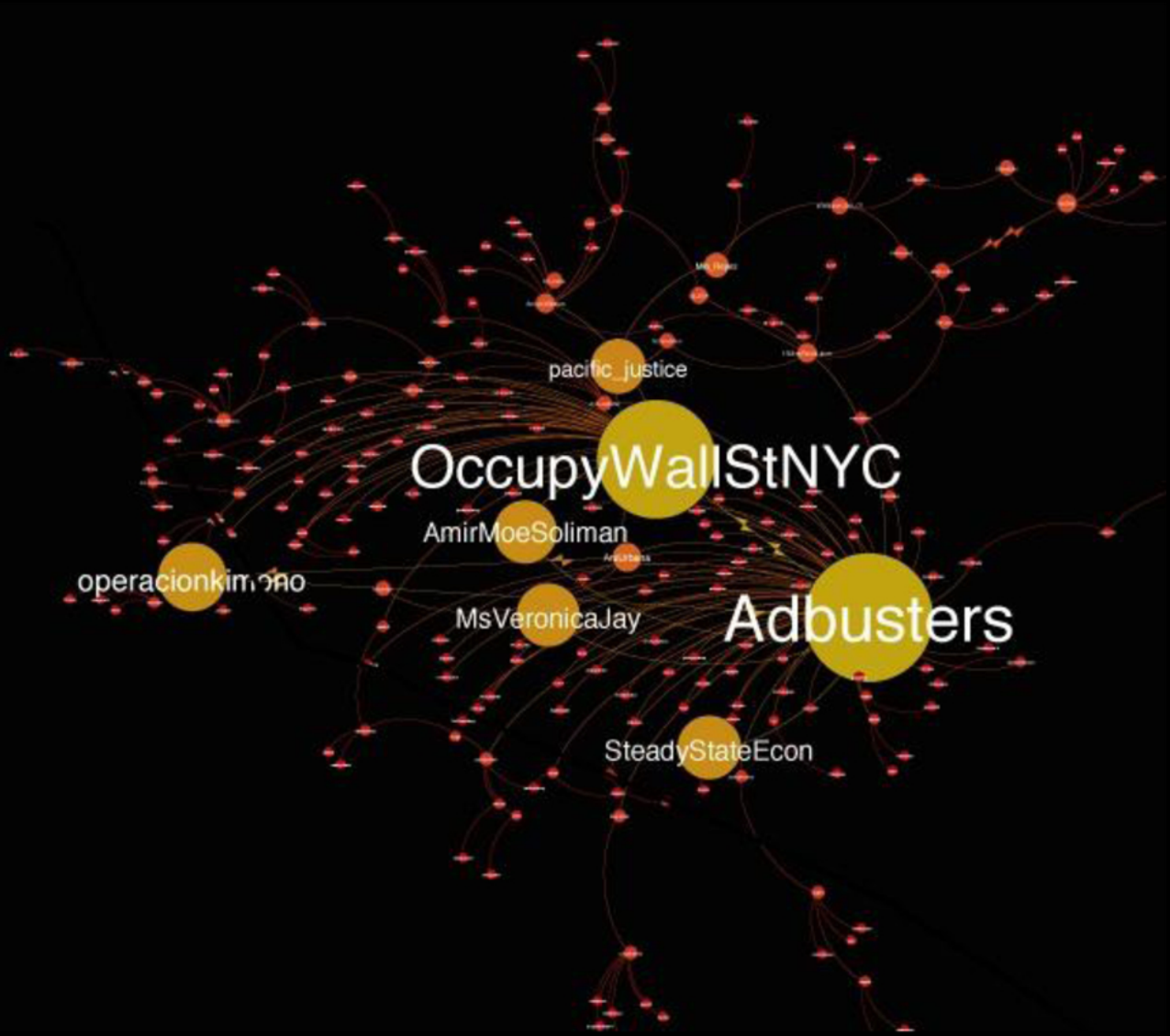


Distance?

PA Graph



Importance of Individual Nodes



Important Metrics

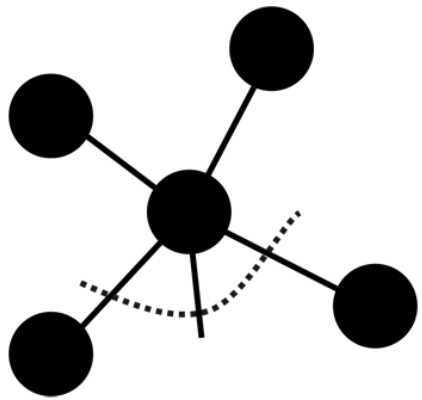
- *Degree Distribution*
- *Clustering Coefficient*
- *Mean Geodesic Distance*
- *Degree Correlation Coefficient*

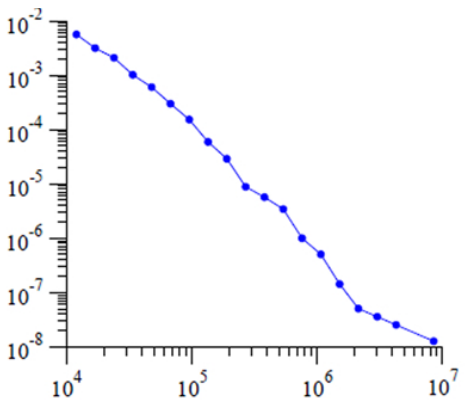
Degree Distribution

The simplest metric we can imagine is the degree of a node, *the number of edges connected to it*.

- Despite its simplicity it is a very successful metric:
  - Indicates the importance of individual nodes.
  - Describes the pattern of connections on an aggregate level (intuitive picture of the network).
  - Analytically tractable.
- Most real world networks exhibit a power-law degree distribution :

$$p_k = C \cdot k^{-a}$$





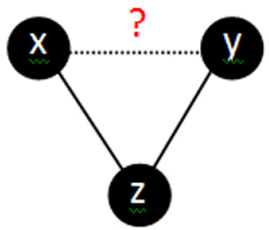
Transitivity

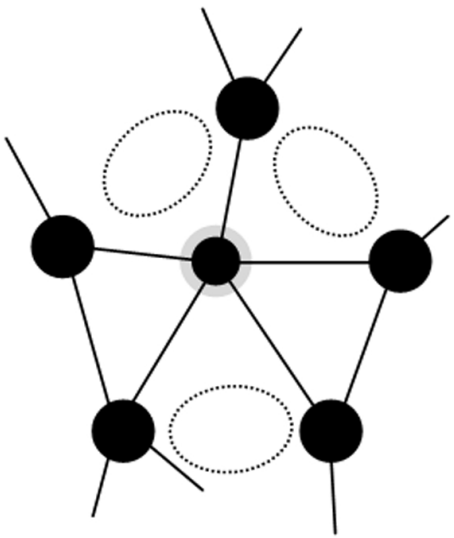
Given that **x** *and* **y** *are two friends of* **z**, *what is the probability that* **x** *and* **y** *are friends?*

- Clustering Coefficient quantifies the degree of transitivity:

$$C = \frac{(number\ of\ triangles \times 3)}{(number\ of\ connected\ triples}$$

- If $C = 0$ we have no closed triangles (e.g tree, lattice), whereas if $C = 1$ we have only cliques.





Network	Film Actors	Biology	Email
Real	0.20	0.09	0.16
Random	0.0003	0.00001	0.00002

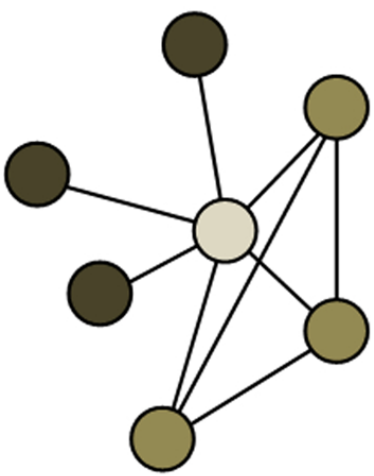
Closeness Centrality

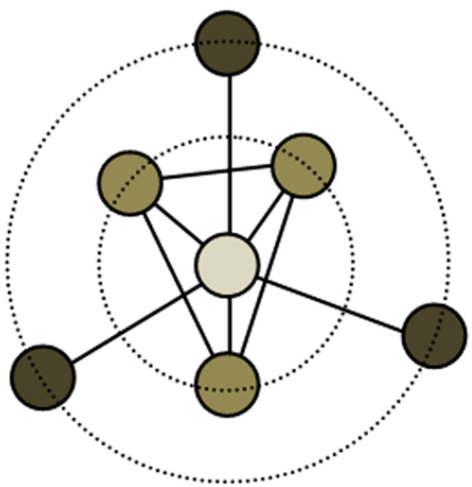
How close is a given vertex to any other vertex in the network on average?

$$C_i = \frac{1}{\ell_i} = \frac{n}{\sum_j d_{i,j}}$$

where ℓ_i is the mean geodesic distance of node i. This metric is closely related to the study of the "small world phenomenon" and its success lies in:

- Main phenomena rely on shortest paths(transportation, cascades, epidemics)
- Provides clear "functional" ordering of the centrality of nodes.





Assortative Mixing

In many real world networks, especially in social networks, there is a tendency for people to be friends with people with similar characteristics (nationality, age, job, hobby). This phenomenon is called *homophily*. We call such networks *assortative*.

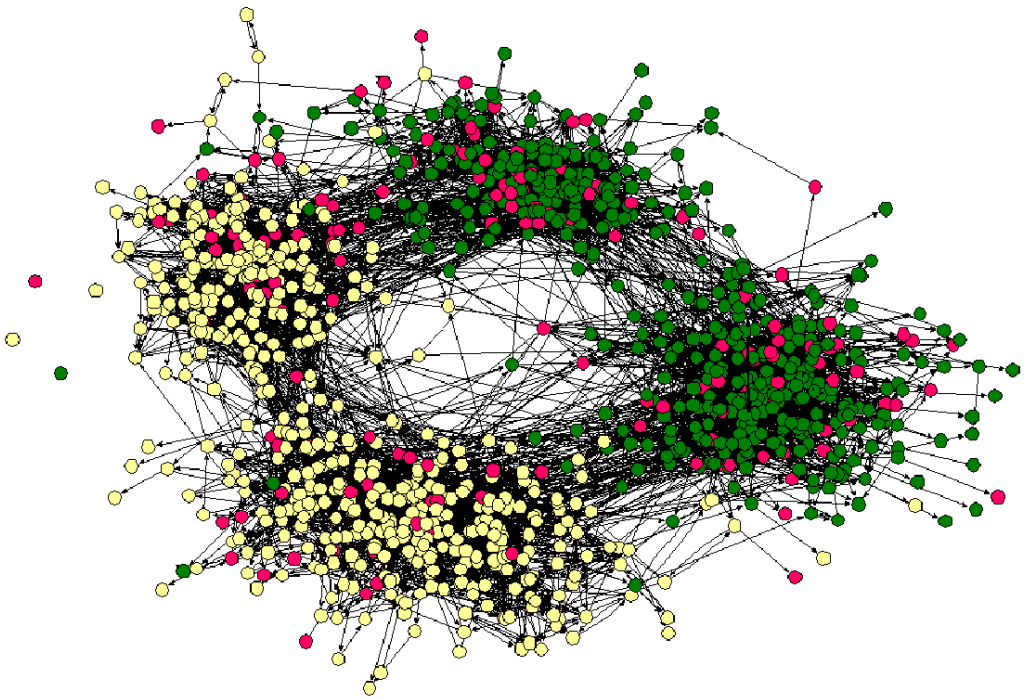


Figure: Friendship relations in "Countryside High School" by race and gender. James Moody, *Race, school integration, and friendship segregation in America, American Journal of Sociology 107*

Degree Correlation Coefficient

- We would like to have a meaningful measure of the assortativity of a network. Assume we have for each node i a scalar characteristic x_i, then a natural metric would be to consider the covariance of x_i and x_j over all edges:

$$cov(x_i, x_j) = \frac{\sum_{ij} A_{ij}(x_i - \mu)(x_j - \mu)}{\sum_{ij} A_{ij}} = \frac{1}{2m} \sum_{ij} \left(A_{ij} - \frac{k_i k_j}{2m} \right) x_i x_j$$

  where $\mu = \frac{1}{2m} \sum_i k_i x_i$ is the mean value of x_i at the end of an edge.

- We normalize the above measure by dividing with the value that a perfectly mixed network would have, $\frac{1}{2m} \sum_{ij} \left(k_i \delta_{ij} - \frac{k_i k_j}{2m} \right) x_i x_j$. So, if we want to find the correlation between degrees we set $x_i = k_i$ and obtain the *Degree Correlation Coefficient*:

$$r = \frac{\sum_{ij} (A_{ij} - \frac{k_i k_j}{2m}) k_i k_j}{\sum_{ij} \left(k_i \delta_{ij} - \frac{k_i k_j}{2m} \right) k_i k_j}$$

Assortative vs. Disassortative

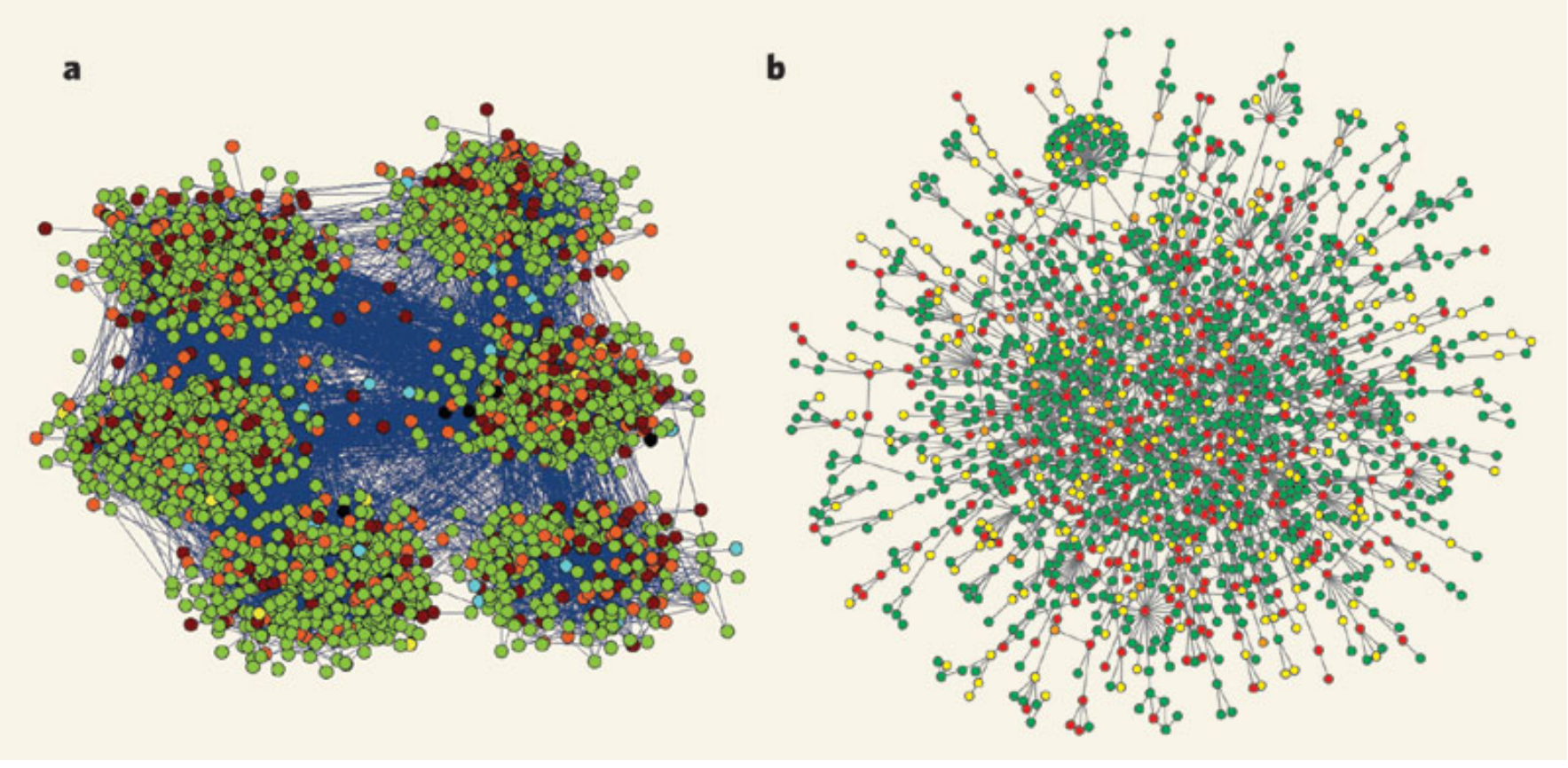
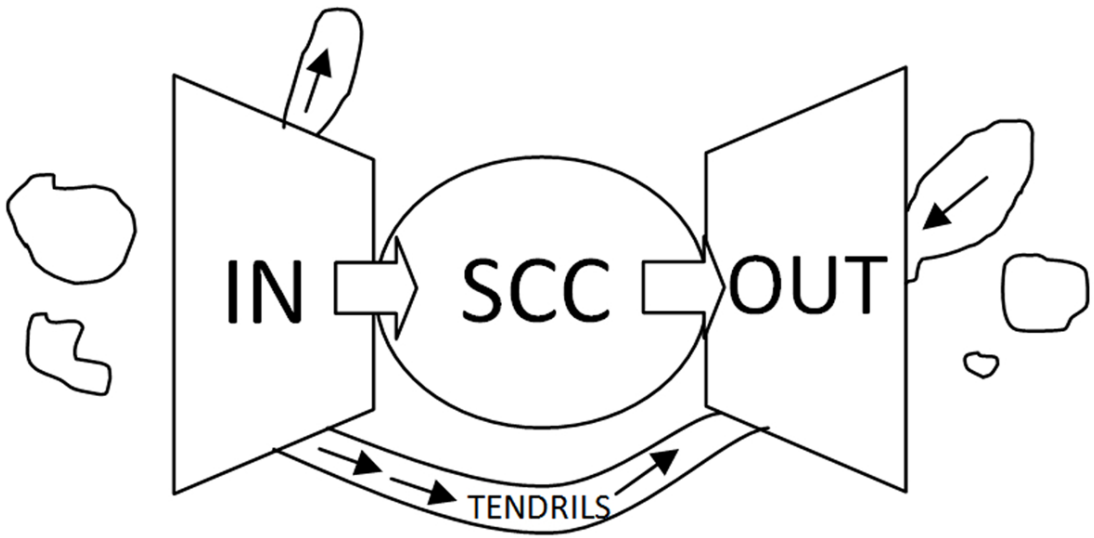


Figure: **a,** In assortative networks high degree nodes tend to stick together so we expect a structure with a dense core and more sparse peripheries. **b,** In disassortative networks, by contrast, well-connected nodes join to a much larger number of less-well-connected nodes creating star-like structure. Figure from *Networks: Teasing out the missing links, Sid Redner, Nature 453*

Structure of the Internet

- The component structure of directed networks is more complicated than for undirected ones. We have weakly and strongly connected components. The equivalent of the giant component in undirected networks is the Giant **Strongly Connected Component** and it is associated with an in- and out-component.



- In directed networks there is an in- and out-degree distribution as well. It has been found that both distributions follow a **power law with exponents** 2.1 **and** 2.7 respectively.

Statistics of Real World Networks

Network	Nodes	Edges	c	S	Distance	Exponent	C	r
Film Actors	449 913	25 516 482	113.43	0.98	3.48	2.3	**0.2**	0.208
Dating	573	477	1.66	0.5	16.01	-	**0.005**	-0.029
WWW	203 million	1466 million	7.2	0.914	16.18	2.1/2.7	**-**	-
Citation	783 369	6 716 198	8.57	-	-	3.0/-	**-**	-
Internet	10 697	31 992	5.98	1.00	3.31	2.5	**0.035**	-0.189
Metabolic	765	3 686	9.64	0.996	2.56	2.2	**0.090**	-0.240
Protein	2 115	2 240	2.12	0.689	6.80	2.4	**0.072**	-0.156
Neural	307	2 359	7.68	0.967	3.97	-	**0.18**	-0.226

In general we can observe that many real world networks exhibit:

- *power law degree distributions*,
- *small world phenomenon*,
- *clustering and degree correlation*
- *assortativity*

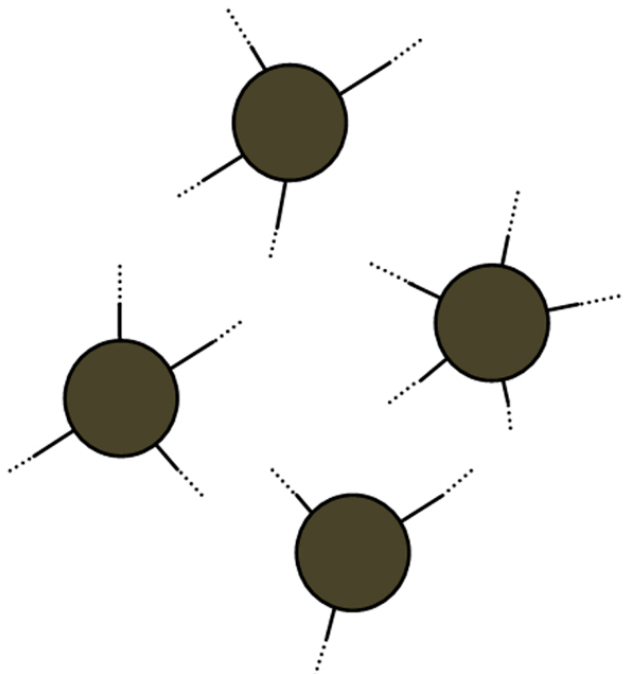
A accurate network model should incorporate all of these phenomena. A first step would be to have a model that realizes any degree distribution.

Configuration Model

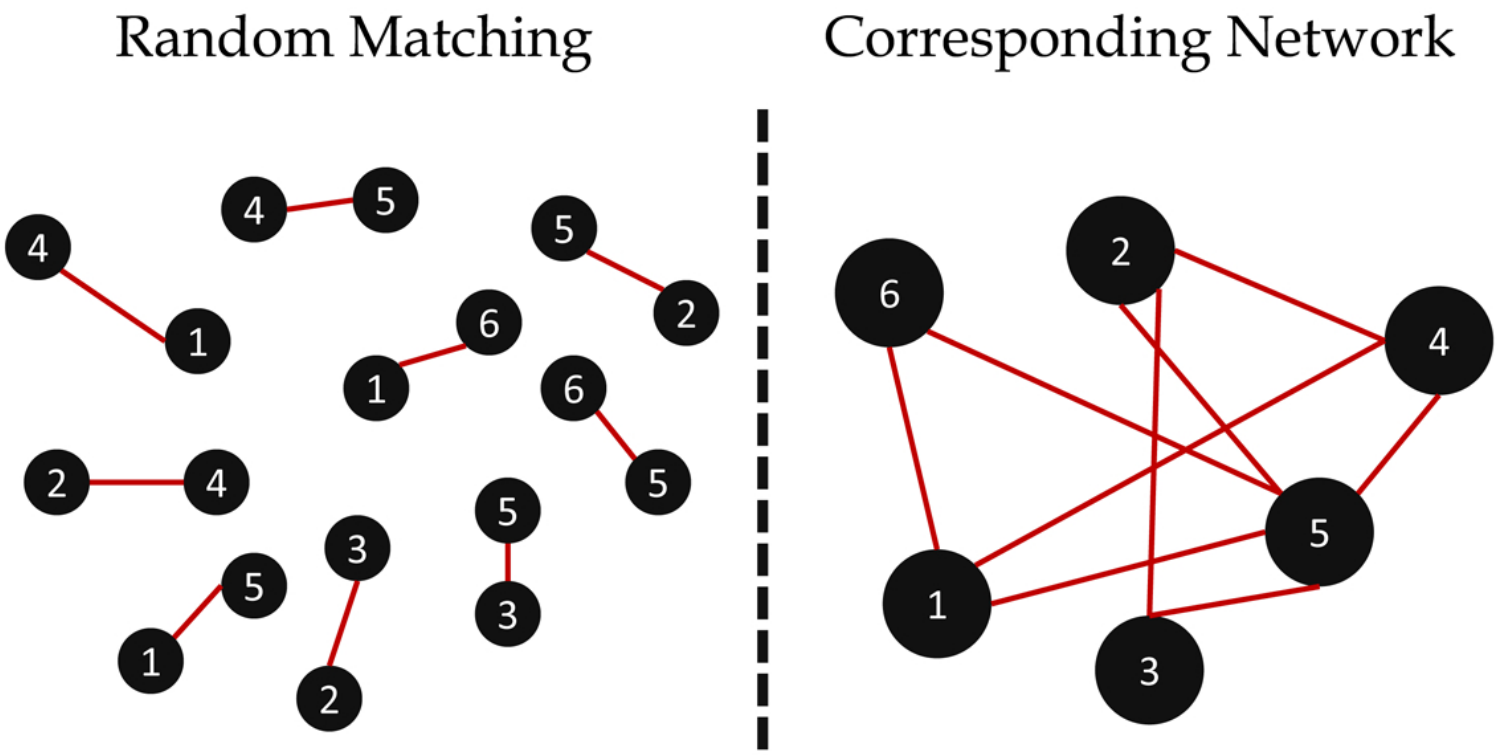
Configuration Model

The *Configuration Model* is actually a model of a random graph with a given degree sequence instead of distribution. In this model we are given the degree k_i of each vertex. This model is analogous to $G_{n,m}$ as the number of edges is fixed.

- For every vertex with degree k_i, we create k_i open ended edges (*stubs*).
- The graph is created by *iteratively selecting uniformly at random two "stubs" and connecting them*.
- In other words we select a random matching.



Configuration Model



Self-Edges, Multi-edges and other Evils

There are a number of issues with the configuration model that gives us trouble:

- The network may contain self-edges or multi-edges or both. However, the **density of self edges tends to zero** in the limit.
- Unfortunately, not all networks with the given degree distribution appear with the same probability as due to self/multi-edges some networks appear more often.

Edge Probability

- The probability p_{ij} of the occurrence of an edge between two specific vertices i and j.

$$p_{ij} = \frac{k_i \cdot k_j}{2m-1} = \frac{k_i \cdot k_j}{2m}$$

- Once we have an edge between vertices i and j then the probability of a second edge is $(k_i - 1)(k_j - 1)/2m$. The expected number of multi-edges is:

$$\frac{1}{2}\sum_{ij}\frac{k_i k_j}{2m}\cdot\frac{(k_i-1)(k_j-1)}{2m} = \frac{1}{2}\left[\frac{\langle k^2\rangle - \langle k\rangle}{\langle k\rangle}\right]^2$$

Excess Degree

- Consider a configuration model with degree distribution p_k. What is the probability that *if we choose a vertex and random and follow one of it's edges we find a vertex with degree k*?

$$\frac{k}{2m} \times np_k = \frac{kp_k}{\langle k \rangle}$$

- We will be interested in what follows with the number of edges attached to a vertex other than the edge we arrived along. This number is called the *excess degree*. We can calculate the fraction of vertices with excess degree k simply by noting that these vertices must have degree $k+1$, hence:

$$q_k = \frac{(k+1)p_{k+1}}{\langle k \rangle}$$

Your Friends are more Popular than you

Consider a randomly chosen vertex and the average degree $\langle d \rangle$ of its neighbours:

$$\langle d \rangle = \sum_k k \frac{k p_k}{\langle k \rangle} = \frac{\langle k^2 \rangle}{\langle k \rangle} > \langle k \rangle$$

- If we were modelling a friendship network then that would mean that on average *"your friends have more friends than you do"*. Actually , if one performs these calculations on real networks the result still holds.
- The reason for this counter-intuitive fact is that a vertex of degree k will be present in k averages. This means that high degree vertices are over-represented and it is this bias that pushes the value up.

Clustering Coefficient

Recall that the **Clustering Coefficient** is the average probability that two neighbours of a vertex are neighbours themselves.

- Consider a vertex v with two neighbours i and j with excess degrees k_i and k_j. They are connected with probability $k_i k_j / 2m$, and if we sum over all the possible degrees we get:

$$C = \sum_{k_i, k_j = 0}^{\infty} q_{k_i} q_{k_j} \frac{k_i k_j}{2m} = \frac{1}{2m} \left[\sum_{k=0} k q_k \right]^2 = \frac{1}{n} \frac{[\langle k^2 \rangle - \langle k \rangle]^2}{\langle k \rangle^3}$$

- This expression decreases as n^{-1} and so vanishes in the limit of large n like in the Poisson Random Graph. However, for some distributions the second moment $\langle k^2 \rangle$ diverges, so we find surprisingly large values of C for the configuration model.

Generating Functions

- if X is a discrete random variable taking values in the non-negative integers $\{0, 1, \ldots\}$ then the probability-generating function of X is defined as:

$$g(z) = \mathbb{E}\left(z^X\right) = \sum_{k=0}^{\infty} p_k z^k = p_0 + p_1 z + p_2 z^2 \ldots$$

  where p_k is the *probability mass function* of X.

- The generating function encapsulates all of the information about the probability distribution in a single function, since:

$$p_k = \frac{1}{k!} \frac{d^k g(0)}{dz^k}$$

$$\langle k^m \rangle = \left[\left(z \frac{d}{dz} \right)^m g(z) \right]_{z=1}$$

Powers of Generating Functions

- Consider a distribution p_k and the corresponding generating function $g(z)$. Suppose, we have m integers which are independently drawn from this distribution. Let π_s be the probability that the integers add up to s, then:

$$\pi_s = \sum_{k_1=0}^{\infty} \cdots \sum_{k_m=0}^{\infty} \delta(s, \sum_i k_i) \prod p_{k_i}$$

- We want to calculate the generating function $h(z)$ of the sum of those integers, which in this case are degrees:

$$\begin{aligned} h(z) &= \sum_{s=0}^{\infty} \pi_s z^s = \sum_{s=0}^{\infty} z^s \left(\sum_{k_1=0}^{\infty} \cdots \sum_{k_m=0}^{\infty} \delta(s, \sum_i k_i) \prod p_{k_i} \right) \\ &= \sum_{k_1=0}^{\infty} \cdots \sum_{k_m=0}^{\infty} \prod p_{k_i} z^{k_i} = \left[\sum_{k=0}^{\infty} p_k z^k \right]^m = [g(z)]^m \end{aligned}$$

Generating Functions for Degree Distributions

- In order to further study the properties of the Configuration Model we will need the generating functions of the degree distribution and the excess degree distribution, which are respectively:

$$g_0(z) = \sum_{k=0}^{\infty} p_k z^k \quad , \quad g_1(z) = \sum_{k=0}^{\infty} q_k z^k$$

- The above generating functions are not really independent, since the excess degree distribution is defined in terms of the ordinary degree distribution:

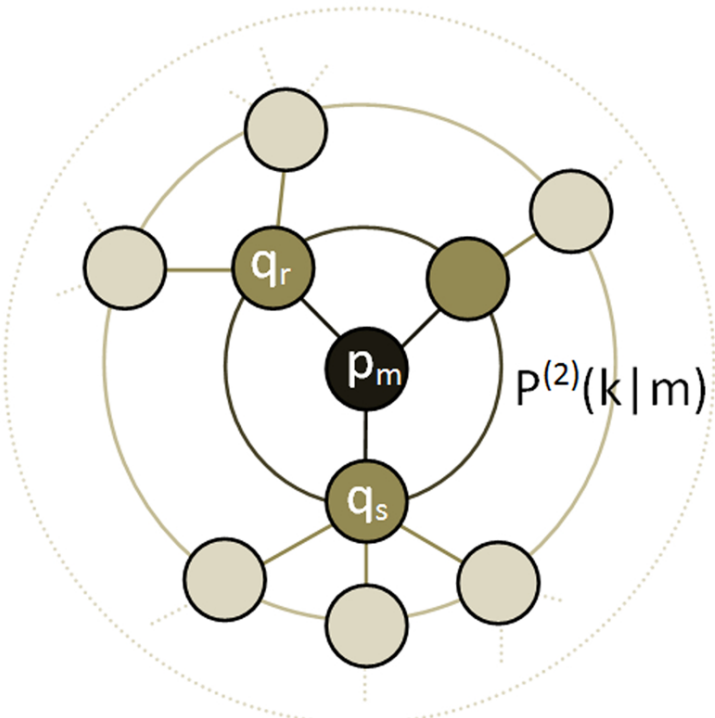
$$g_1(z) = \frac{1}{\langle k \rangle} \sum_{k=0}^{\infty} (k+1) p_{k+1} z^k = \frac{1}{\langle k \rangle} \sum_{k=0} k p_k z^{k-1} = \frac{1}{\langle k \rangle} \frac{dg_0}{dz}$$

Number of Second Neighbours

To understand the component structure of the Configuration model, we will need to calculate the distribution of number of second neighbours of a randomly chosen vertex $p^{(2)}(k) = \sum_{m=0}^{\infty} p_m P^{(2)}(k|m)$, where $\mathbf{P^{(2)}(k|m)}$ is the probability that the sum of second neighbours of a vertex are k given that he has m first neighbours.

- We can write the generating function for the number of second neighbours:

$$\begin{aligned} g^{(2)}(z) &= \sum_{k=0}^{\infty} \left(\sum_{m=0}^{\infty} p_m P^{(2)}(k|m) \right) z^k \\ &= \sum_{m=0}^{\infty} p_m \sum_{k=0}^{\infty} P^{(2)}(k|m) z^k \\ &= \sum_{m=0}^{\infty} p_m [g_1(z)]^m \\ &= g_0(g_1(z)) \end{aligned}$$



Neighbours at distance d

- Similarly, we can show that the number of neighbours at any (small) distance d has generating function:

$$g^{(d)}(z) = \sum_{m=0}^{\infty} p_m^{(d-1)} [g_1(z)]^m = g^{(d-1)}(g_1(z))$$

- We can use this equation to calculate the average number c_d of neighbours at distance d, which for any distribution is given by $\frac{dg^{(d)}}{dz}$ for $z = 1$:

$$c_d = g^{(d-1)'}(1) g_1^{'}(1) = c_{d-1} g_1^{'}(1)$$

- Since $c_1 = g_0^{'}(1) = \langle k \rangle$ and that $c_2 = g_0^{'}(1) g_1^{'}(1)$, we get:

$$g_1^{'}(1) = \frac{\langle k^2 \rangle - \langle k \rangle}{\langle k \rangle} = \frac{c_2}{c_1}$$

Giant Component

- If we solve the previous recurrence we get that the mean number of nodes at distance d is:

$$c_d = \left(\frac{c_2}{c_1}\right)^{d-1} c_1$$

- That means that once we know c_1 and c_2, we know everything. The average number of nodes at distance d either grows or falls exponentially, depending on the ratio c_2/c_1. As a consequence, we have the condition for the appearance of the giant component:

$$c_2 > c_1 \Leftrightarrow \langle k^2 \rangle - 2\langle k \rangle > 0 \Leftrightarrow \sum_i k_i(k_i - 2) > 0$$

- Observe that the condition implies that nodes with degree 0 and 2 don't matter for the existence of the giant component. We must note that the above reasoning is not exact, as there are cycles and multi-edges. However, the result holds and was shown first in 1995 by Molloy and Reed by using entirely different techniques.

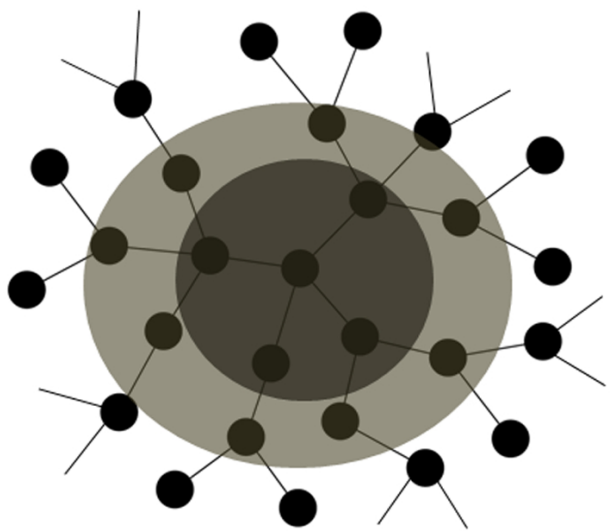
Molloy and Reed '95

Explore the graph by BFS process. The probability that we pick a vertex of degree i to explore is proportional to its degree. The expected increase is roughly:

$$\mathbf{Q(D) = \sum_{i \geq 1} i(i-2)d_i(n)}$$

If $Q(D) < 0$ the BFS process will die out. Whereas, if $Q(D) > 0$ then the component might grow quite large.

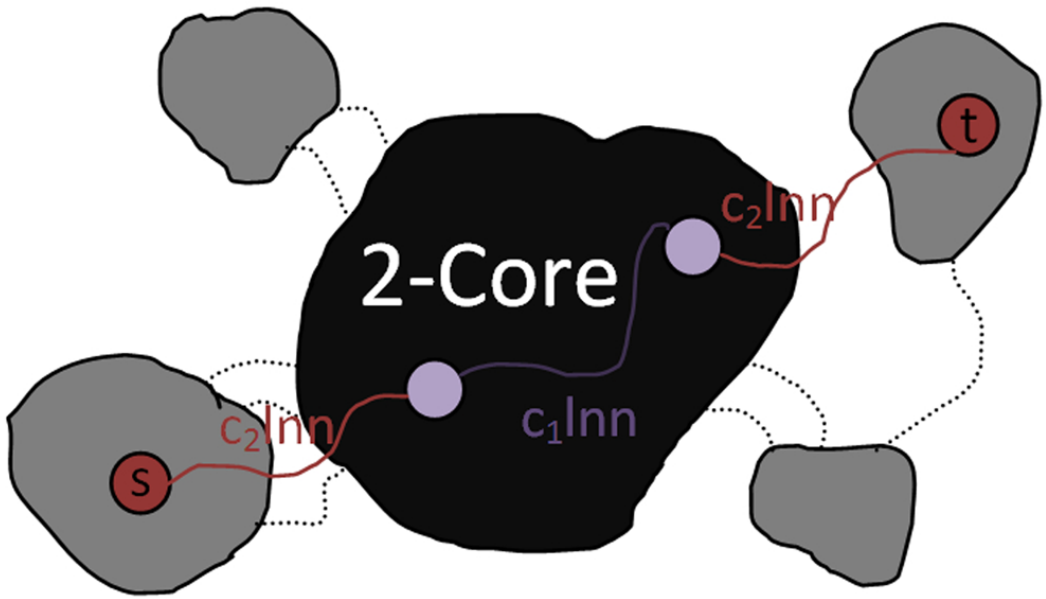
- Principle of Deferred Decisions
- "De-dimensionalize" by projecting the process on a unit dimensional Random Walk.
- Use coupling to deal with irregularities
- Concentration results by employing Method of Bounding Differences and Branching Processes Techniques



The Diameter

Fernholz and Ramachandran in 2007 proved that if there is a giant component then the longest distance in the graph is:

$$D = c \cdot lnn = (2c_2 + c_1)lnn$$



In their proof they considered a **BFS procedure restricted only on the 2-core** of the graph and managed to trace the rate of neighbourhood expansion. They showed that the longest path would consist of a path form a vertex s to the $2-$core of length $c_2 lnn$, a path in the $2-$core of length $c_1 lnn$ and a path from the $2-$core to the other vertex t of length $c_2 lnn$.

Other Results

- [Gkantsidis, Mihail, Saberi '03]
  Congestion of routing is $O(n \cdot log(n))$(expanders) for power law graphs.
- [Frieze, Krivevelich, Smyth '06]
  The Chromatic Number is $\Theta(\frac{d}{logd})$ where d is the average degree. For some family of distributions(e.g. power laws and exponential distribution).
- [Van der Hoefstadt+coauthors'04'05'08]
  Studied extensively diameter and typical distances for power law distributions $x^{-\alpha}$.
  - $(\alpha > 2)$ $\Theta(log(N))$.
  - $(2 < \alpha < 3)$ $\Theta(log(N))$ if min $deg. \geq 3$ then $O(log(log(N)))$.
  - $(1 < \alpha < 2)$ Constant.
- [Cooper, Frieze, Krivevelich '10]
  Hamiltonicity for CM when power laws or exponential distribution.
- [Bohman, Picollelli '11]
  SIR epidemics solved by method of differential equations[Wormald].

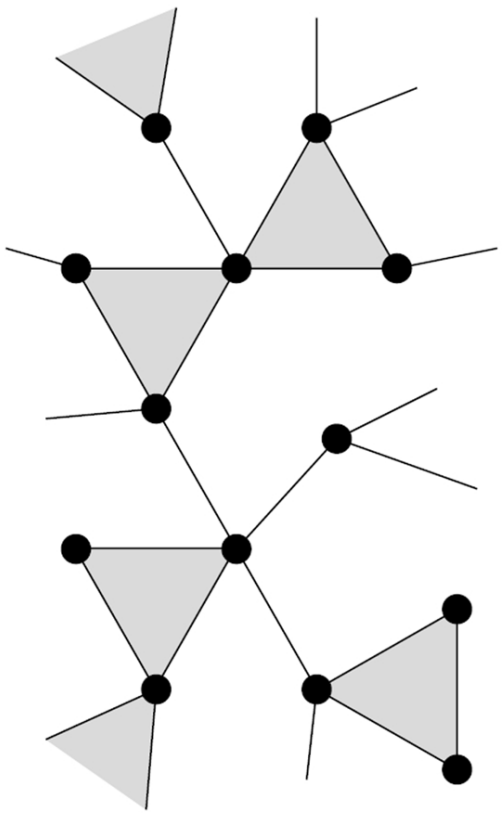
Networks with Transitivity

The idea of the *Configuration Model* can be generalized further and include more complicated features. The idea is to specify the number of other subgraphs as well, and then consider "hyperedges" by selecting the vertices that will participate to the subgraph. We present here a model from *Random Graphs with Clustering by M.E.J. Newman, 2009:*

- In this model we specify both the number of edges and the number of triangles. Let t_i and s_i be the number of triangles and single edges that vertex i participates. We define the joint degree distribution p_{st} with generating function:

$$g_p(x, y) = \sum_{s,t=0}^{\infty} p_{st} x^s y^t$$

- We can think of each vertex having a number of "stubs" and "corners". The network is created by selecting a random matching independently for stubs and corners.



Networks with Transitivity

- If we know the joint distribution $g_p(x, y)$, the distribution for the total degree distribution can be found easily:

$$f(z) = \sum_{k=0}^{\infty} p_k z^k = \sum_{k=0}^{\infty} \sum_{s,t=0}^{\infty} p_{s,t} z^{s+2t} = g_p(z, z^2)$$

- We can use these generating functions to calculate the number of triangles and the number of connected triples:

$$\begin{aligned} 3N_{\triangle} &= n \sum_{st} t p_{st} = n \left(\frac{\partial g_p}{\partial y} \right)_{x=y=1} \\ N_3 &= n \sum_{k} \binom{k}{2} p_k = \frac{1}{2} n \left(\frac{\partial^2 f}{\partial z^2} \right)_{z=1} \end{aligned}$$

- Thus, we wil always have a non zero value for the **clustering coefficient** in the limit of large n:

$$C = \frac{3N_{\triangle}}{N_3} = 2 \left(\frac{\partial g_p}{\partial y} \Big/ \frac{\partial^2 f}{\partial z^2} \right)_{x=y=z=1}$$

Configuration Model Variant

- Given the similarity of the Configuration Model to $G_{n,m}$, it is natural to assume if there is also a $G_{n,p}$ equivalent for this case as well. We can define a similar model by selecting a parameter c_i for each vertex and placing each edge independently with probability:

$$p_{ij} = \begin{cases} \frac{c_i c_j}{2m} & for i \neq j \\ \frac{c_i^2}{4m} & for\ i = j \end{cases}$$

- With this choice the average degree of vertex i is:

$$\langle k_i \rangle = 2p_{ii} + \sum_{j \neq i} \frac{c_i c_j}{2m} = \sum_{ji} \frac{c_i c_j}{2m} = c_i.$$

- In other words the parameters c_i are just the expected degrees. Thus, in this model we don't have a fixed degree distribution and the degrees are random variables. In fact one can show that the degree of vertex i will be poisson distributed with mean c_i. **This variation in the degree sequence distribution makes the model unsatisfactory for modelling networks**.

Evaluation

The Configuration Model consists a **firm step** towards modelling real world networks as:

- It can materialize any valid degree sequence both for directed and undirected networks.
- It exhibits the small world phenomenon.
- With proper adjustments it can exhibit high degrees of clustering.
- It can be generalized to include any number of fixed subgraphs and gives us great modelling capabilities.

Evaluation

However, there are a number of unpleasant facts that makes it **unsatisfactory**:

- It allows multiedges and hyper-multiedges, which is unrealistic.
- In order to make it a realistic we must to resort to overfitting, specifying a number of parameter distributions (degree, clustering, correlations).
- It provides no insight in how these networks might have actually formed through temporal evolution and its mechanism has no real life counterpart.

Preferential Attachment

Preferential Attachment

- Networks such as the Internet, citation networks are observed to approximately follow power law degree distributions. The power law is somewhat an unusual distribution and its appearance is an indicator of an interesting underlying process. *How do these power laws appear?*

Rich get Richer, Feedback, Preferential Attachment

- Preferential Attachment is a mechanism that gives an advantage/bias to more prosperous nodes making them morel likely to be selected in the future by other nodes.

Origins

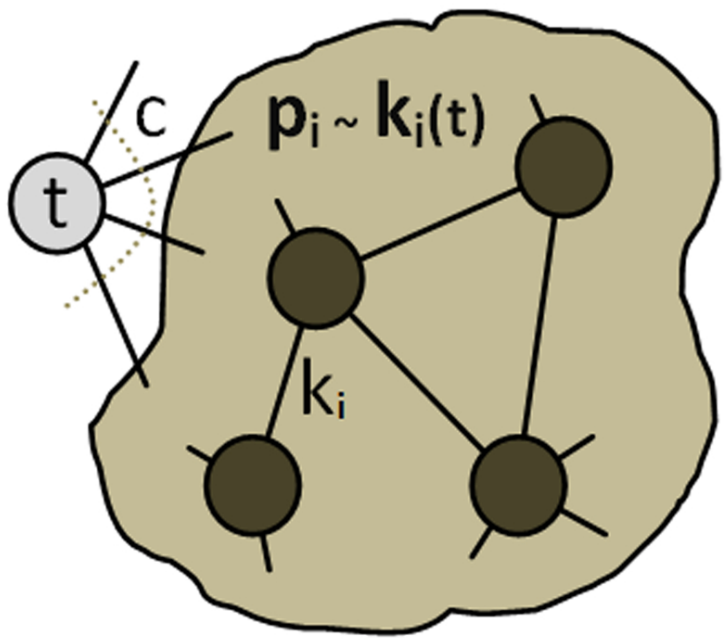
- **Price had studied in the 1960's the citation network** of scientific papers and pointed the power law degree distribution. In the 1970's Price considered the question of how these power laws appear and proposed a simple and elegant model of network formation.
- In **1999 Barabasi and Albert**, unaware of Price's work, considered a similar mechanism and showed that it results in power law distributions. They coined the mechanism as *Preferential Attachment*.
- **Krapivsky and Redner in 2000**, extended the *Preferential Attachment* mechanism for non-linear kernels as well. Subsequently Bianconi and Barabasi extended the mechanism for non-uniform kernels where each node had a different attractiveness.

Price's Model

Price considered the **citation network growth process**. He assumed that papers(nodes) are published continually and that newly appearing papers cite(edges) previously existed ones. Moreover:

- Every paper sites on average c old papers.
- Each "old" paper is cited with probability proportional to the number of citations q_i it already has plus a constant a. Thus:

$$p_i = \frac{q_i + a}{\sum_i (q_i + a} = \frac{q_i + a}{n(c + a)}$$



Master Equation Approach

- Each new paper cites c vertices on average, thus a vertex with degree q receives on average $c\frac{q+a}{n(c+a)}$ new citations. There are $np_q(n)$ vertices with degree q and they receive on average:

$$np_q(n) \times c\frac{q+a}{n(c+a)} = \frac{c(q+a)}{(c+a)}p_q(n)$$

- From the population of vertices with degree q we gain 1 for each vertex of degree $q-1$ that receives a new citation from the new node, and lose 1 for each vertex of degree q:

$$gain = \frac{c(q-1+a)}{(c+a)}p_{q-1}(n) \quad loss = \frac{c(q+a)}{(c+a)}p_q(n)$$

- Now we can write the *master equation* for the evolution of the in-degrees as follows:

$$(n+1)p_q(n) = np_q(n) + \frac{c(q-1+a)}{(c+a)}p_{q-1}(n) - \frac{c(q+a)}{(c+a)}p_q(n)$$

Solving the Recurrence

- Now let us consider the limit of large network size and calculate the asymptotic form of the degree distribution. Taking the limit $n \to \infty$ and setting $p_q = p_q(\text{inf})$, we get the recurrence:

$$p_q = \frac{c}{c+a}[(q-1+a)p_{q-1} - (q+a)p_q] \ , \ p_0 = \frac{1+a/c}{a+1+a/c}$$

- If we solve for p_q we get:

$$p_q = \frac{q+a-1}{q+a+1+a/c}p_{q-1} \ , \ p_0 = \frac{1+a/c}{a+1+a/c}$$

- If we recursively substitute the lower terms we have:

$$p_q = \frac{(q+a-1)(q+a-2)\dots a}{(q+a+1+a/c)(a+2+a/c)} \frac{1+a/c}{a+1+a/c}$$

Asymptotics

- Luckily, the solution for p_q can be expressed in closed form due to the properties of the $\Gamma(x)$ and $B(x, y)$ functions:

$$\frac{\Gamma(x+n)}{\Gamma(x)} = (x+n-1)(x+n-2)\dots x \ , \ B(x,y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}$$

- After some algebraic manipulations we have the solution in closed form:

$$p_q = \frac{B(q+a, 2+a/c)}{B(a, 1+a/c)}$$

- If we use standard Stirling approximations for $\Gamma(x)$ we derive asymptotic approximations for $B(x, y)$:

$$B(x,y) \simeq \frac{e^{-x}x^{x-\frac{1}{2}}}{e^{-(x+y)}x^{x+y-\frac{1}{2}}e^{y}} = x^{-y}\Gamma(y)$$

Power Laws!

- Applying the asymptotic calculation in the closed form solution for the degree distribution we get:

$$\mathbf{p_q} \sim (\mathbf{q} + \mathbf{a})^{-\alpha} \sim \mathbf{q}^{-\alpha}$$

- Where the exponent a is given by: $\alpha = 2 + \frac{a}{c}$. Thus, Price's model give rise to power law distributions with exponent between $2 < a < \infty$. Barabasi and Alberts's model is just a special case where $a = c$ and results in a power law with exponent 3.
- Note also that the condition of Molloy and Reed implies that there is a giant component if :

$$\langle k^2 \rangle - 2\langle k \rangle > 0 \Rightarrow \zeta(\alpha - 2) > 2\zeta(\alpha - 1) \Rightarrow \alpha < 3.4788$$

Temporal Evolution

- Intuitively **Price's model gives an advantage to vertices created earlier**, as they have more time to acquire links from other vertices. We want to get a quantitative estimate for this fact. Let $p_q(t, n)$ be the fraction of vertices created at time t that have in-degree q when the network has n vertices. We can write the master equation:

$$(n+1)p_q(t, n+1) = np_q(t, n) + \frac{c}{c+a}[(q-1+a)p_{q-1}(t, n) - (q+a)p_q(t, n)].$$

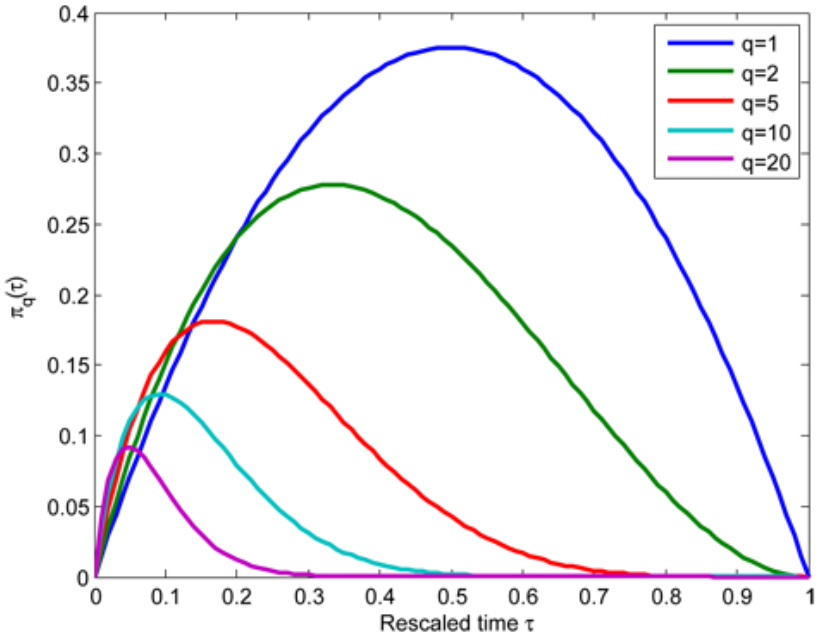
- Because at each time t only one vertex is created, it would make more sense to change to a rescaled time: $\tau = \frac{t}{n}$. In the same time we also change the density from $p_q(t, n)$ to $\pi_q(\tau, n) = np_q(t, n)$, the fraction of vertices with degree q created at times form τ to $\tau + d\tau$. The master equation becomes for $n \to \infty,\ \epsilon = 1/n$:

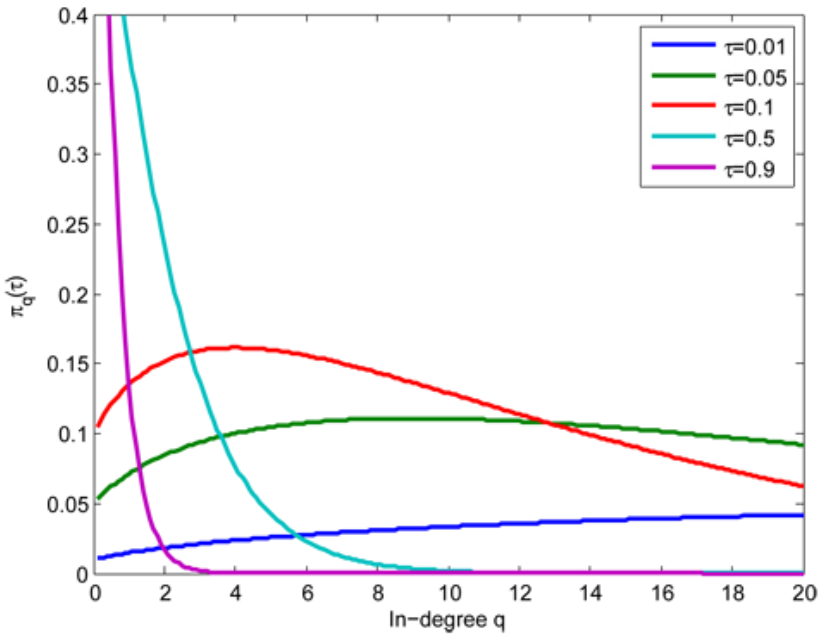
$$\frac{\pi_q(\tau) - \pi_q(\tau - \epsilon\tau)}{\epsilon} + \frac{c}{c+a}[(q-1+a)\pi_{q-1}(\tau) - (q+a)\pi_q(\tau)] = 0$$

- As $n \to \infty,\ \epsilon \to 0$ the master equation becomes a differential equation:

$$\tau\frac{d\pi_q}{d\tau} + \frac{c}{c+a}[(q-1+a)\pi_{q-1}(\tau) - (q+a)\pi_q(\tau)] = 0$$

Temporal Evolution





- If we solve for lower degrees to higher starting from $\pi_0(\tau)$, we get that:

$$\pi_q(\tau) = \frac{\Gamma(q+a)}{\Gamma(q+1)\Gamma(a)} \tau^{ca/(c+a)} (1 - \tau^{ca/(c+a)})^q$$

- For large q we have that $\pi_q(\tau)$ decays exponentially except for a leading algebraic factor:

$$\pi_q(\tau) \sim q^{a-1} (1 - \tau^{ca/(c+a)})^q$$

First Mover Advantage

- We see that vertices added to the network early have a significant advantage over those added later. In the setting of citations this suggest that early papers in a field will receive substantially more citations than later ones, purely because they were published first.
- This is the **fundamental feedback mechanism** that appears in many situations. A small initial lead is amplified by preferential attachment process and turns into a significant advantage.
- However, this is a crude model as it ignores the quality of individual papers as well as the "hotness" of a field. We thus need to generalize the PA model, to more complex *attachment kernels*.

Non-Linear PA

- Studies of real world networks have shown that the growth rate isn't exactly linear. Rather it has been found that it follows a power law: $a_k = k^\gamma$ with varying exponents γ. For $\gamma = 1$ we have linear preferential attachment as before. These are data from *Measuring Preferential Attachment for evolving networks, Jeong et. al.*

Network	Nodes	Edges	γ
Citation	*1736*	*83252*	***0.95 ±0.1***
Internet	*12409*	*13445*	***1.05***
Collaboration	*209293*	*3534724*	***0.79 ±0.1***
Actor	*392340*	*33646882*	***0.81 ±0.1***

- What effect would a non-linear preferential attachment have on the degree distribution of a network? Let a_k be the functional form of the kernel. Then the correct normalized probability would be:

$$\Pi(k) = \frac{a_k}{\sum_i a_{k_i}}$$

Non-Linear PA

- Let $\mu(n) = \frac{1}{n}\sum_i a_{k_i} = \sum_k a_k p_k(n)$ then we can write the master equation as before:

$$(n+1)p_k(n+1) = np_k(n) + \frac{c}{\mu(n)}[a_{k-1}p_{k-1}(n) - a_k p_k(n)].$$

- Taking the limit for $n \to \infty$ and writing $p_k = p_k(\infty),\ \mu = \mu(\infty)$, we have :

$$p_k = \frac{c}{\mu}[a_{k-1}p_{k-1} - a_k p_k],\ p_c = \frac{\mu/c}{a_c + \mu/c}$$

- If we solve the recurrence we get that:

$$p_k = \frac{\mu}{ca_k}\prod_{r=c}^{k}\left[1 + \frac{\mu}{ca_r}\right]^{-1}$$

Power Law Kernel

If we consider the power law form of the attachment kernel where $a_k = k^\gamma$ and use asymptotic analysis we find that:

- In the sublinear regime, where $\gamma < 1$ a power law degree distribution with exponential cutoff occurs.

$$p_k = k^{-\gamma} exp\left[-f(\mu, \gamma)\right]$$

- In the linear case for $\gamma = 1$, we have a tunable exponent power law degree distribution, as describes previously.
- In the super-linear regime $\gamma > 1$, a condensation to gel-like state occurs, where a "winner takes all" phenomenon occurs and a single node is connected to allmost all nodes in the network and the power law disappears.

Fitness Models

- One property of the PA models so far is the conservation of hubs over time. In this setting a latecomer node will have a significant disadvantage in obtaining a high degree. However in real world networks some new nodes may surpass older nodes due to their *internal fitness*.
- **Bianconi and Barabasi in 2001** proposed a model that assigns to every node i inserted in the network a fitness η_i drawn from a distribution $\rho(\eta)$. The probability that a newly formed node will connect to an existing node j is :

$$\Pi_i = \frac{\eta_j k_j}{\sum_\ell \eta_\ell k_\ell}$$

- Writing again the master equation and taking the limit $n \to \infty$, we get:

$$p_k(\eta) = \frac{c}{\mu}[a_{k-1}(\eta)p_{k-1}(\eta) - a_k(\eta)p_k(\eta)]$$

Degree Distribution

- Using the form we derived for non-linear PA with the attachment kernel $a_k(\eta) = \eta \cdot k$ and the initial condition $p_c(\eta) = \rho(\eta) - \frac{ca_c(\eta)}{\mu} p_c(\eta)$, we get that:

$$p_k(\eta) = \rho(\eta) \frac{B(k, 1 + \mu/c\eta)}{B(c, \mu/c\eta)} \sim \rho(\eta) k^{1+\frac{\mu}{c\eta}}$$

- We have shown that the distribution of degrees of vertices with a particular value of fitness follows a power law, but what about the overall distribution? Let us first calculate the mean degree $\langle k \rangle$:

$$\langle k \rangle = \int_0^\infty \sum_{k=c}^{\infty} k p_k(\eta) d\eta = c \int_0^\infty \frac{\rho(\eta) d\eta}{1 - c\eta/\mu}$$

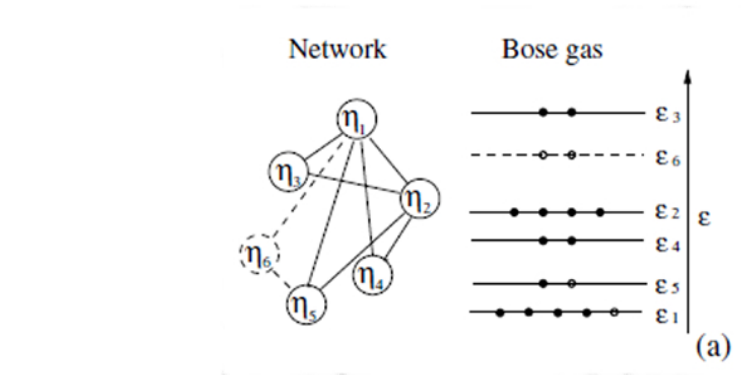
Condensation

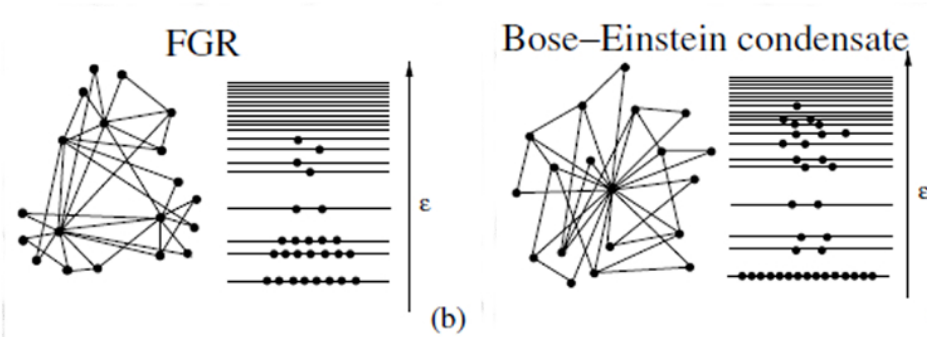
Bianconi and Barabasi solved the model utilizing a mapping from networks to Bose-Einstein gases. For each node they introduced an energy level and for each edge they introduce 2 particles in the system. The degrees in the original network are equivalent to the number of particles at each energy level. They found that:

(a) When all the fitness values are the same we have the *scale free phase* and we have pure power law.

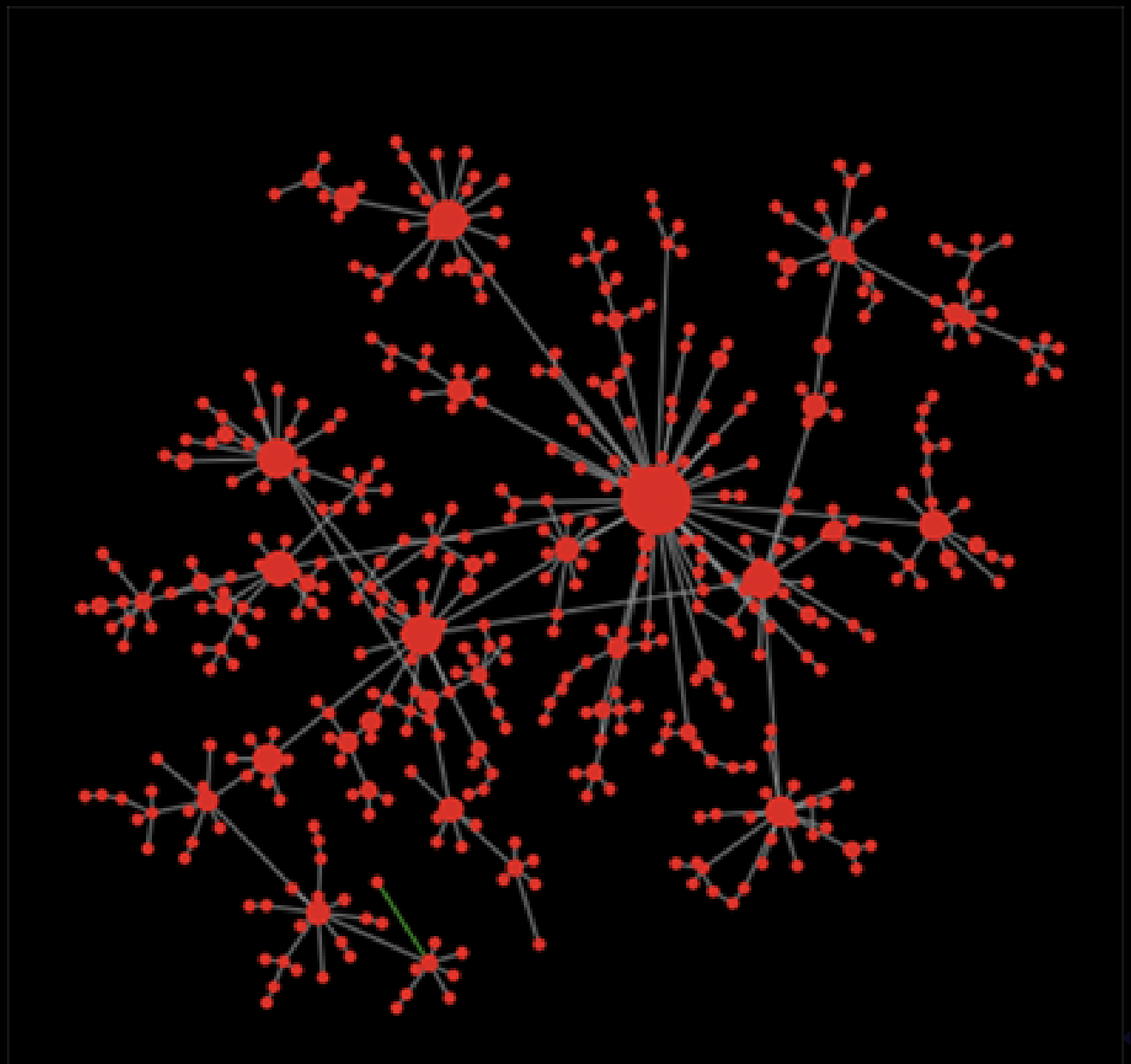
(b) If we have and arbitrary fitness distribution, the high fitness nodes become the hubs. This is the *Fit-Get-Richer phase*.

(c) If $\int_0^{\eta_0} \frac{\rho(\eta)d\eta}{1-c\eta/\mu} < 2$ then we have *"condensation"* where one or more vertices accumulate a finite fraction of the edges.





Preferential Attachment Network



Other Results

- [Bollobas, Riordan '00]
  BA model($\tau = 3$) diameter is $\Theta(log(N))$ if $c = 1$, and $\Theta(\frac{log(N)}{log(log(N))})$.
- [Flaxman, Frieze, Vera '06'08]
  Geometric Preferential Attachment on a sphere in $\mathbb{R}^3$.
- [Souza, Borgs, Chayes, Berger, R.Kleinberg '06]
  Emergence of Preferential Attachment from optimization.
- [Borgs, Chayes, Daskalakis, Roch STOC'07]
  Rigorous analysis of PA with Fitness via Generalized Polya Urn model. Confirmed Bianconi-Barabasi predictions and obtained analytic results.
- [Fronczak, Fronczak, Holyst '08] Clustering in BA model is roughly: $C(t) = \frac{(c-1)log(t)^2}{8t}$
- [Dommers, Van der Hoefstadt, Hooghiemstra '09]
  Diameter is $\Theta(log(t))$ for $\tau > 3$ and $O(log(log(t)))$ for network with $2 < \tau < 3$.

Evaluation

The **Preferential Attachment** models are quite adequate as models of real world networks, since:

- Power law degree distribution emerge quite naturally.
- They are small worlds with distances scaling as $log(n)$ for $\alpha > 3$ and as $\log\log(n)$ for $\alpha \in (2, 3)$.
- Provide intuition in the generative process of the network.
- Exhibit robustness/resilience in random failures/attacks.

- There are also many extensions of these models, for instance one can remove randomly edges, rewire edges or even remove nodes. Surprisingly, many of these cases can be examined analytically through the formalism of generative functions.
- Although, these models don't exhibit high degree of clustering they are up to now the most widely accepted models for Networks and are used extensively in modelling Social Networks, the Web and in Biological settings.

Exponential Random Graphs

Ensembles

- Many of the networks we study exist in only one instantiation and if their growth process repeated they wouldn't be exactly the same. We are interested not in the exact structure but in the qualitative properties of classes of networks, such us Social Networks, the Internet or Biological Networks.
- We would also want to see how processes such as searching, routing behave not on case-specific instantiations but on a general class of networks.

These considerations have driven us to consider *ensemble models* of networks satisfying some desired properties, meaning a set of possible networks(instantiations) and a probability distribution over them. For this purpose we will use the **exponential random graphs** formalism.

Exponential Random Graphs

In the exponential random graphs formalism we consider a set of possible graphs $\mathcal{G}$ for which:

- We fix the average values of some quantities x_i, e.g degree distribution or number of edges.
- We additionally require that we have a valid probability distribution $P(\mathcal{G})$ over the space of graphs we defined

This problem is under-determined so we need to maximize a criterion. The most natural criterion is the "Gibbs entropy". Now we are ready to write down the formulation:

$$\begin{aligned} \max \quad & S = -\sum_{G \in \mathcal{G}} P(G) ln P(G), \\ s.t \quad & \sum_{G \in \mathcal{G}} P(G) = 1 \\ & \sum_{G \in \mathcal{G}} P(G) x_i(G), \ i = 1, \dots \end{aligned}$$

Lagrangian Formulation

- We will solve the optimization problem using the Langrangian Multiplier method, for which we introduce a multiplier for each constraint we must satisfy. If we execute the method we get:

$$P(G) = exp\left[a - 1 + \sum_i \beta_i x_i(G)\right]$$

- Hence the name Exponential Random Graphs. If we set:

$$Z = e^{1-a},\ H(G) = \sum_i \beta_i x_i(G) \Rightarrow Z = \sum_{G \in \mathcal{G}} e^{H(G)}$$

where Z is called the *partition function* and $H(G)$ the *graph Hamiltonian*.

- What remains is to find the values for Z and β_i. These are effectively the parameters of the model and play similar role as p for the Poisson random graph.

Mean Quantities

- Once we have the probability distribution $P(G)$ we can use it calculate estimates of quantities of interest. We are usually interested in mean quantities:

$$\langle y \rangle = \sum_{G \in \mathcal{G}} P(G) y(G) = \frac{1}{Z} \sum_{G \in \mathcal{G}} e^{H(G)} y(G)$$

- We could use this relations for the actual quantities $\langle x_i \rangle$ we want to fix, in order to calculate the unknown quantities β_i:

$$\langle x_i \rangle = \frac{1}{Z} \sum_{G \in \mathcal{G}} e^{\sum_i \beta_i x_i(G)} x_i(G) = \frac{1}{Z} \frac{\partial Z}{\partial \beta_i} = \frac{\partial F}{\partial \beta_i}$$

where $F = lnZ$ is called the *free energy* of the ensemble. Thus, we only need to calculate Z.

Erdos-Renyi Graphs Revisited

Suppose we want a random graph with fixed average number of edges $\langle m \rangle$. Then with the exponential random graph formalism we would have:

$$H = \beta m, \; Z = \sum_G e^{\beta m}.$$

How can we evaluate the sum over all possible graphs. The standard way to achieve this is to sum over possible values of the adjacency matrix A_{ij}, which will be for simple graphs only 0 and 1. Now if we also write $m = \sum_{i<j} A_{ij}$ we get:

$$Z = \sum_{\{A_{ij}\}} \prod_{i<j} e^{\beta A_{ij}} = \prod_{i<j} (1 + e^{\beta}) = \left[1 + e^{\beta}\right]^{\binom{n}{2}}$$

Now using the free energy $F = lnZ = \binom{n}{2} ln(1 + e^{\beta})$ we find that the model is equivalent to $G_{n,p}$:

$$\langle m \rangle = \frac{\partial F}{\partial \beta} = \binom{n}{2} \frac{1}{1 + e^{-\beta}} \Rightarrow \beta = \frac{\langle m \rangle}{\binom{n}{2} - \langle m \rangle}$$

$$p_{ij} = \langle A_{ij} \rangle = \frac{\sum_{A_{ij}=0,1} A_{ij} e^{\beta A_{ij}}}{\sum_{A_{ij}=0,1} e^{\beta A_{ij}}} = \frac{1}{1 + e^{-\beta}} = \frac{\langle m \rangle}{\binom{n}{2}}$$

Fixed Expected Degree Sequence

A natural question is whether we can have an exponential random graph model that has any expected degree sequence. In this model we fix the expected degrees of each vertex. Specifically:

$$H = \sum_i \beta_i k_i, \;\; k_i = \sum_j A_{ij} \Rightarrow H = \sum_{ij} \beta_i A_{ij} = \sum_{i<j} (\beta_i + \beta_j) A_{ij}$$

since we consider symmetric graphs. The partition function now can be written:

$$Z = \sum_{\{A_{ij}\}} exp\left(\sum_{i<j} (\beta_i + \beta_j) A_{ij} \right) = \prod_{i<j} \left[1 + e^{\beta_i + \beta_j} \right]$$

Configuration Model Revisited

We calculate again the probability of an edge between i and j:

$$p_{ij} = \frac{\sum_{A_{ij}=0,1} A_{ij} e^{(\beta_i+\beta_j)A_{ij}}}{\sum_{A_{ij}=0,1} e^{(\beta_i+\beta_j)A_{ij}}} = \frac{1}{1+e^{-(\beta_i+\beta_j)}}$$

Observe that now the edges have different probabilities. Generally we are interested in sparse networks where the probability of any individual edge is small. Thus:

$$p_{ij} \simeq e^{\beta_i} e^{\beta_j} \Rightarrow \langle k_i \rangle = \sum_k p_{ik} = e^{\beta_i} \sum_k e^{\beta_k} = \frac{e^{\beta_i}}{C}$$

If we then sum all the expected degrees $\sum_i \langle k_i \rangle = 2\langle m \rangle$ we get the Configuration Model variant we discussed:

$$p_{ij} = \frac{\langle k_i \rangle \langle k_j \rangle}{2\langle m \rangle}$$

Evaluation

The Exponential Random Graph approach is quite general and natural, however there are some serious difficulties:

- It is very difficult to solve analytically the model even when second or third order constraints are present, e.g triangles, and even if it is possible to have a solution usually we can't create graphs for the full range of the parameters, e.g. transitivity.
- Even if we can't solve it one could argue that we could use MCMC methods and sample from the probability distribution they define. Unfortunately, recently **Bhamidi et.al. (FOCS 2008)** have shown that either the markov chain sampling from the distribution needs exponentially many steps(low temperature) or it mixes really fast(high temperature) but then any finite collection of edges are asymptotically independent making them not really different from Erdos-Renyi graphs.

These results, especially the second, make the use of exponential random graphs extremely limited in the case where analytic results can be found and thus inadequate for modelling as well as simulation purposes.

Other Models

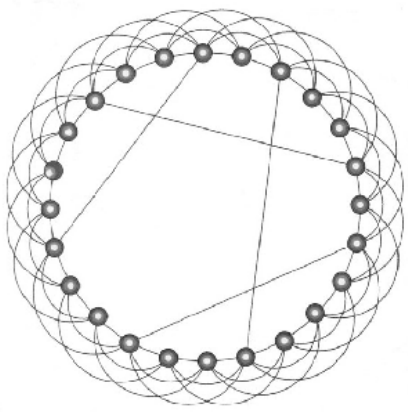


Figure: Lattice Embedded Graphs

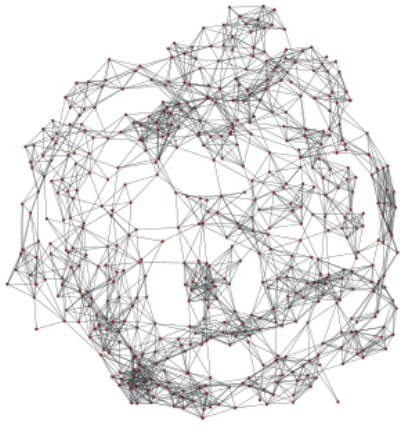


Figure: Geometric Random Graphs

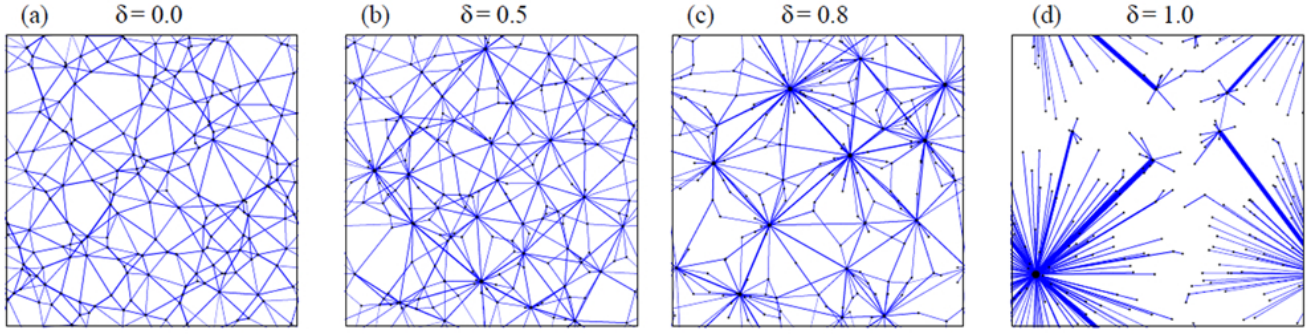


Figure: Network Optimization Models

Discussion

New Frontiers

Since the mid 90's explosion in scientific interest for networks we have come a long way. However, there are still many aspects that are only partially investigated:

- Incorporate Geometry and other "constraints"
- Goal Driven Models and Network Formation Games
- Average case analysis of algorithms and processes

It is a challenge to utilize the knowledge and intuition gained from the study of networks to:

- Design more efficient protocols and procedures(routing, voting, marketing).
- Proactive and intelligent legislation.
- Educate decision-makers to incorporate the underlying network effects in their reasoning.

Harnessing Randomness

Exploring Structure

- Random Graphs and Satisfiability
- PageRank and WebSearch
- Metric Embeddings (Johnson-Lindenstrauss and Achlioptas)
- Linear Predictive Coding

Exploiting Structure

- Randomized Algorithms(Optimization)
- Simulated Annealing

Harnessing Randomness

Exploring Structure

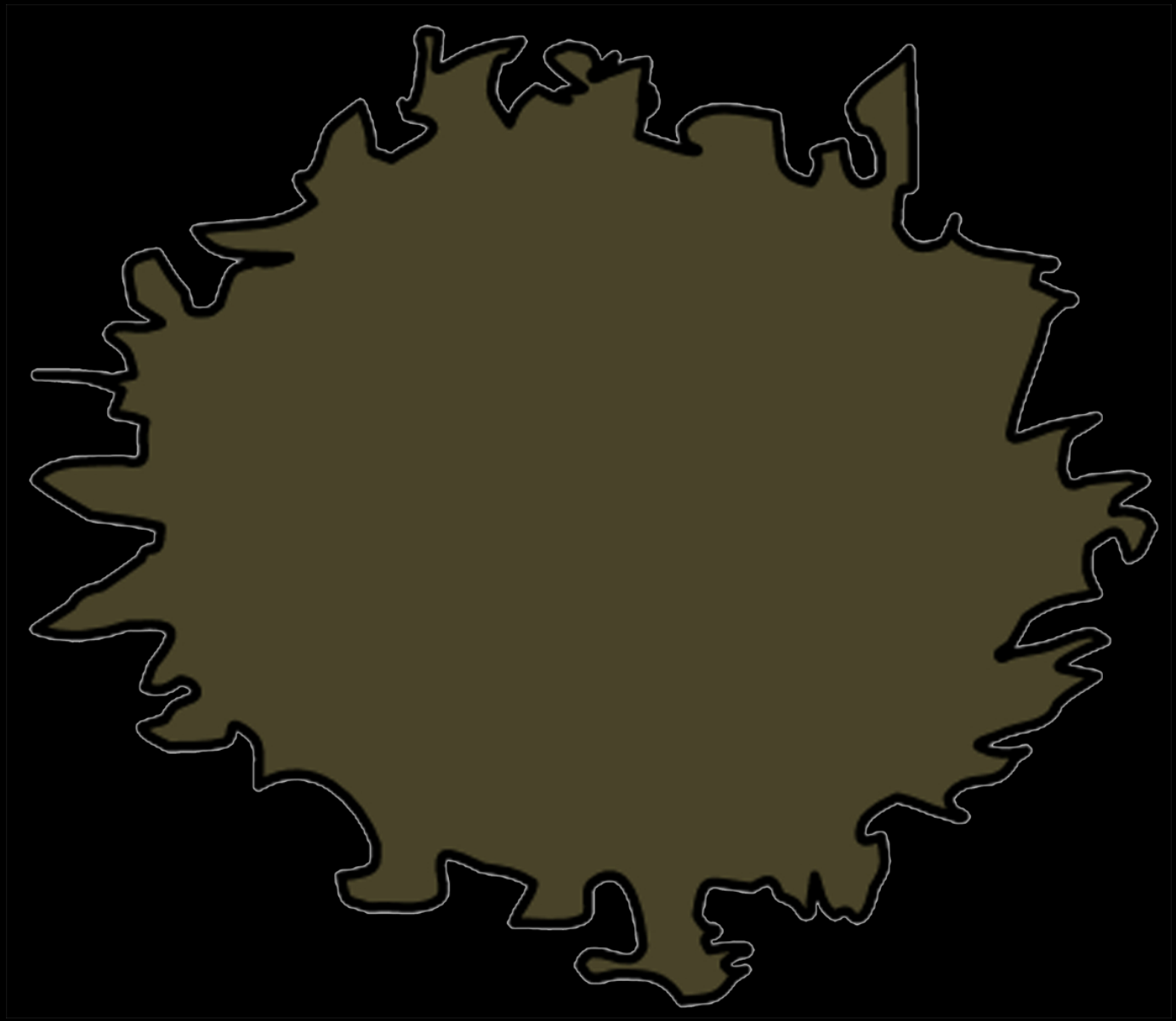
- Random Graphs and Satisfiability
- PageRank and WebSearch
- Metric Embeddings (Johnson-Lindenstrauss and Achlioptas)
- Linear Predictive Coding

Exploiting Structure

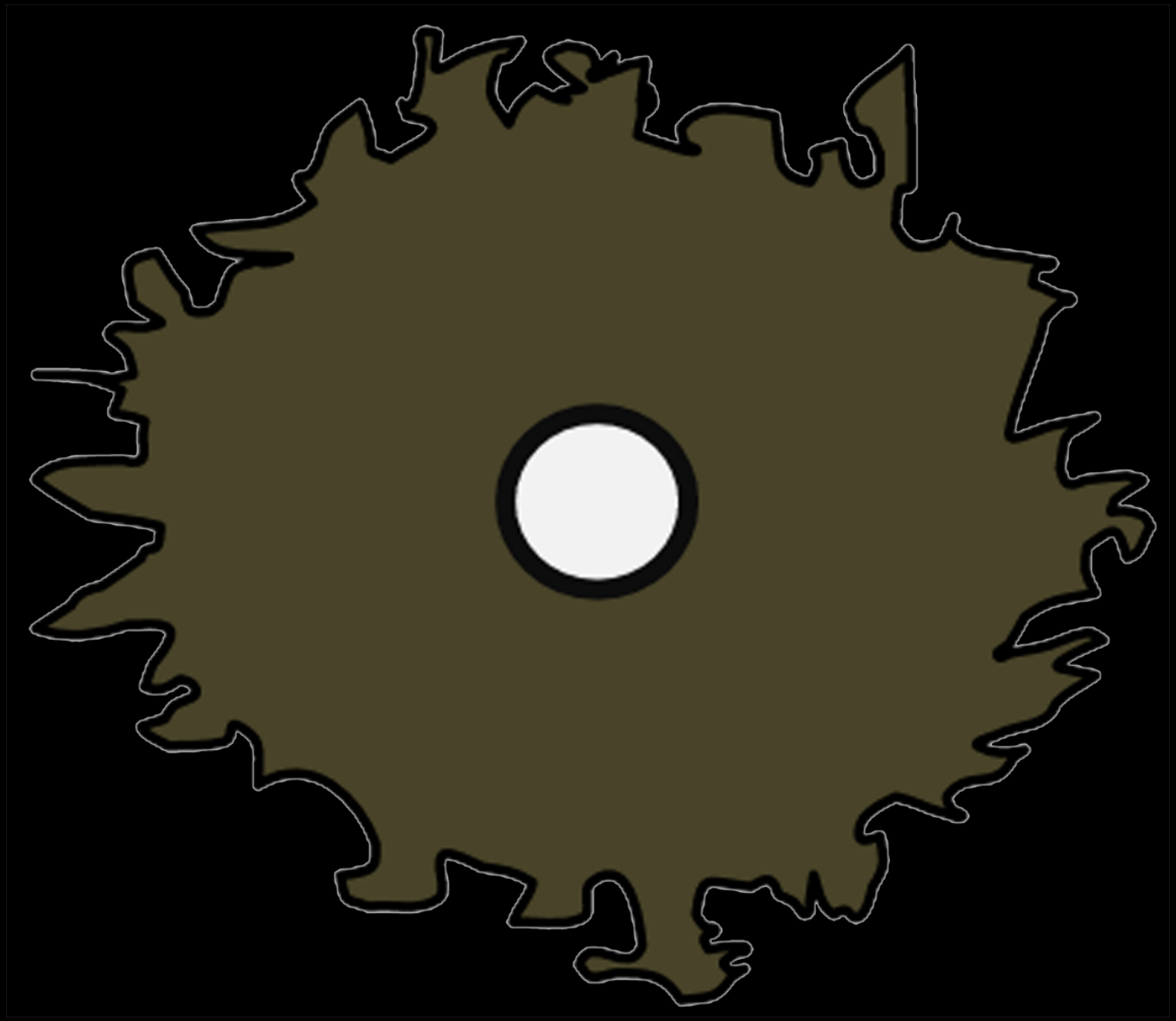
- Randomized Algorithms(Optimization)
- Simulated Annealing

Exploring + **Exploiting** $\rightarrow$ **Evolution**

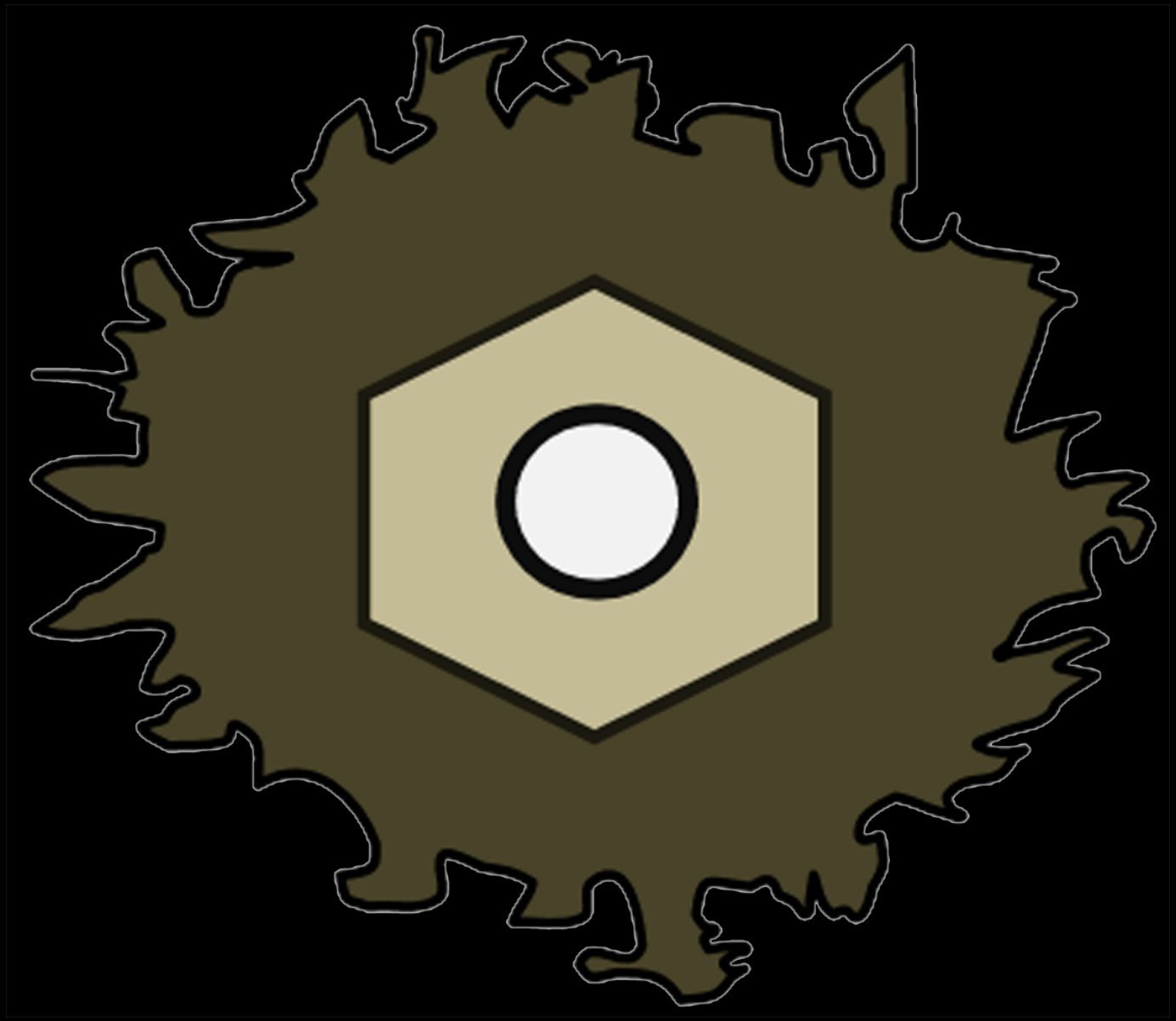
Handling Complexity



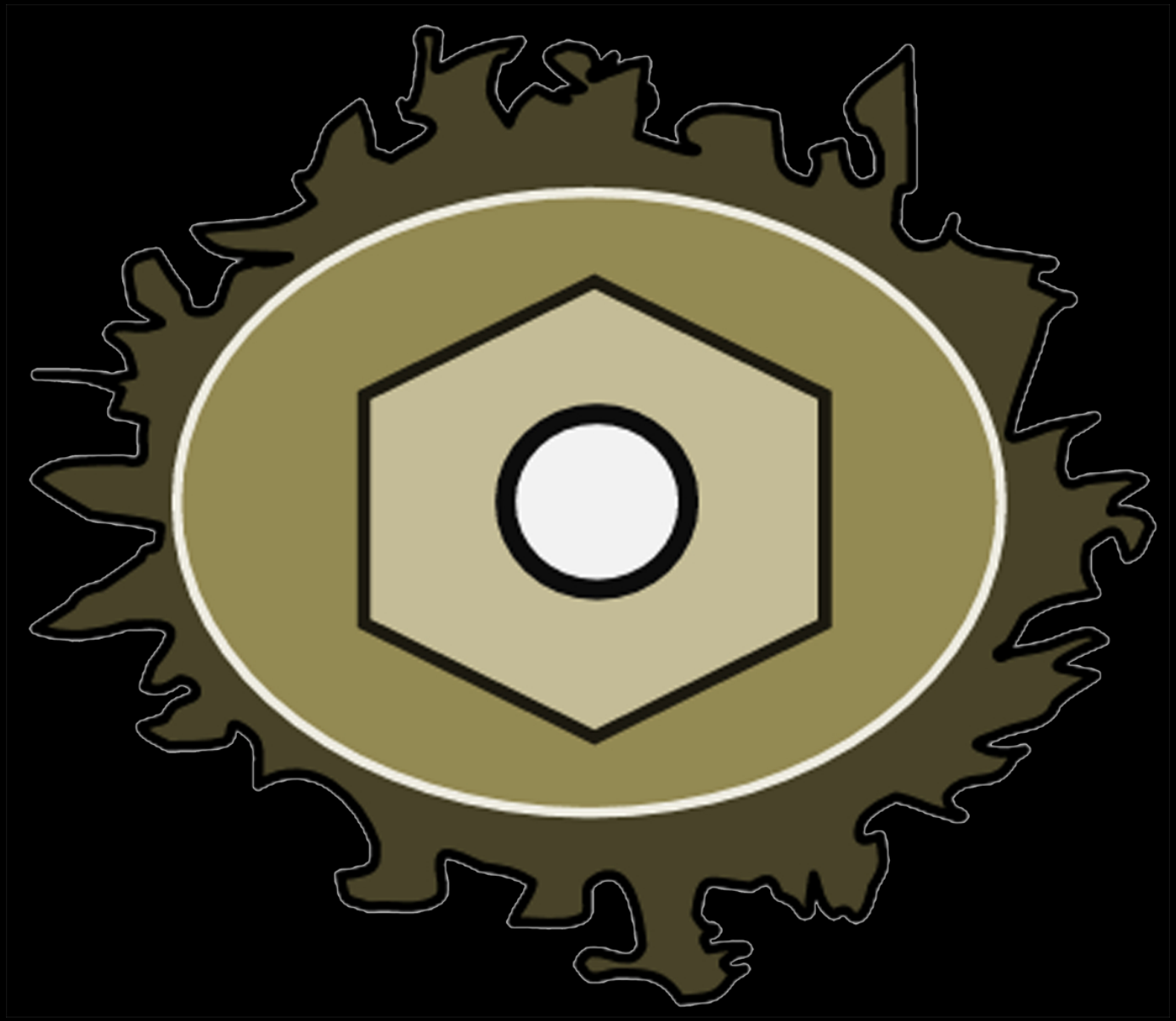
Handling Complexity



Handling Complexity



Handling Complexity



Questions?

Thank You!

psiminelakis@gmail.com

"The point isn't that any oversimplified model will do, but that oversimplified models can go a long way if they get right the few details that really matter. All good science depends on the fact that the important patterns rarely depend sensitively on thousands of factors, but on a crucial few."

Mark Buchanan 2007, *The Social Atom*.

Further Reading

M.E.J Newman
Networks: An Introduction.
Oxford University Press, 2010.

S.N. Dorogovtsev, J.F.F. Mendes
Evolution of Networks
Oxford University Press, 2003.

Reuven Cohen, Shlomo Havlin
Complex Networks: Structure, Robustness and Function
Cambridge University Press, 2010

Mike Molloy, Bruce Reed
A Critical Point for Random Graphs with a Given Degree Sequence
Random Structures and Algorithms 6,pages 161-180 (1995).

S. Bhamidi, G. Bresler, A. Sly
Mixing Time of Exponential Random Graphs
FOCS 2008