Models and Algorithms for Complex Networks

“with network elements, typically local”

“with parametrization maintaining characteristic profiles”

“with categorical attributes”

[C. Faloutsos MMDS08]

Milena Mihail

Georgia Tech.

with

Stephen Young, Gagan Goel, Giorgos Amanatidis, Bradley Green,
Christos Gkantsidis, Amin Saberi,
D. Bader, T. Feder, C. Papadimitriou, P. Tetali, E. Zegura

Talk Outline

Complex Networks in Web N.0

Flexible (further parametrized) Models

1. Structural/Syntactic Flexible Models

2. Semantic Flexible Models

Models & Algorithms Connection : Kleinberg's Model(s) for Navigation

Distributed Searching Algorithms with Additional Local Info/Dynamics

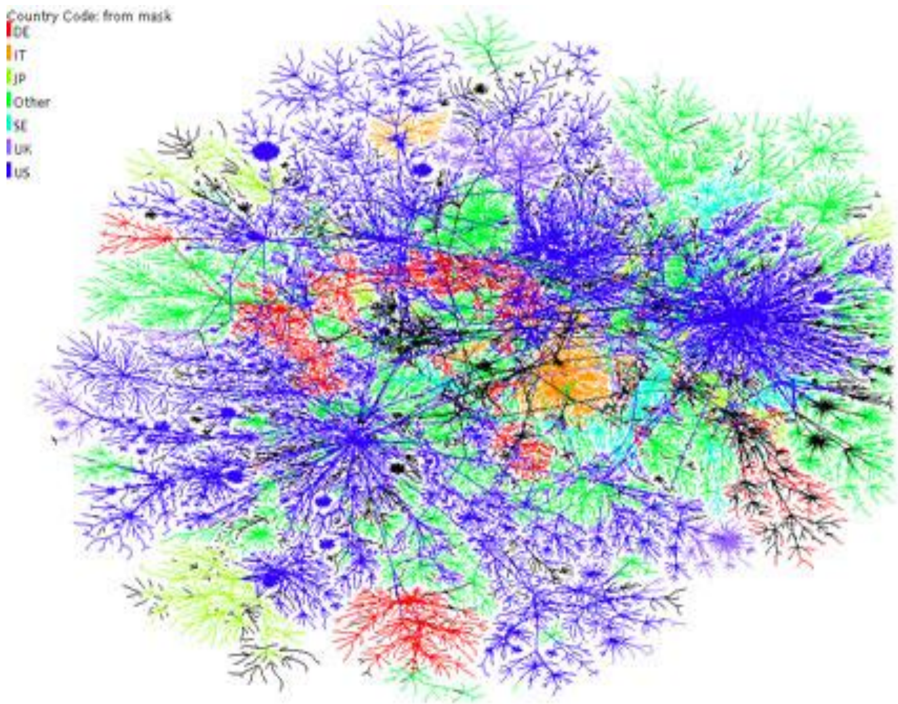
1. On the Power of Local Replication

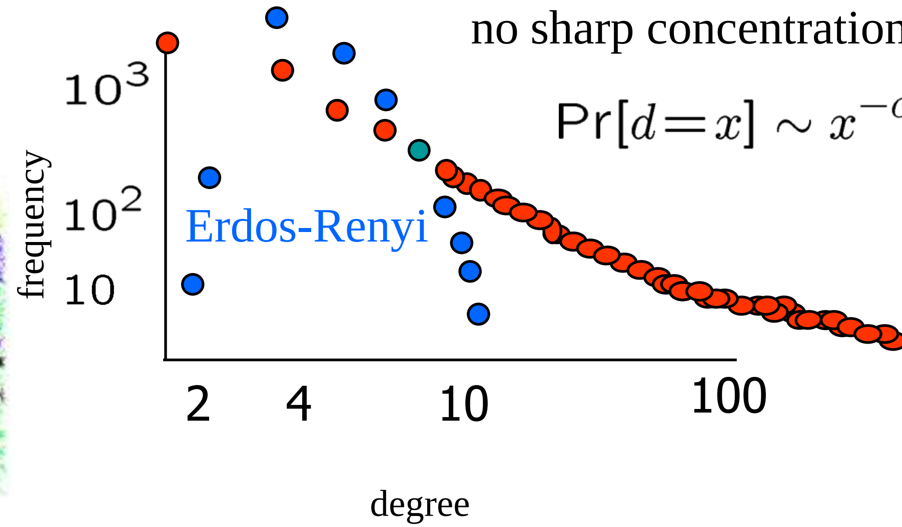
2. On the Power of Topology Awareness via Link Criticality

Conclusion : Web N.0 Model & Algorithm characteristics:
further parametrization, typically local,
locality of info in algorithms & dynamics.
Dynamics become especially important.

$E[d] \simeq 4$, but
no sharp concentration:

$\Pr[d = x] \sim x^{-\alpha}$



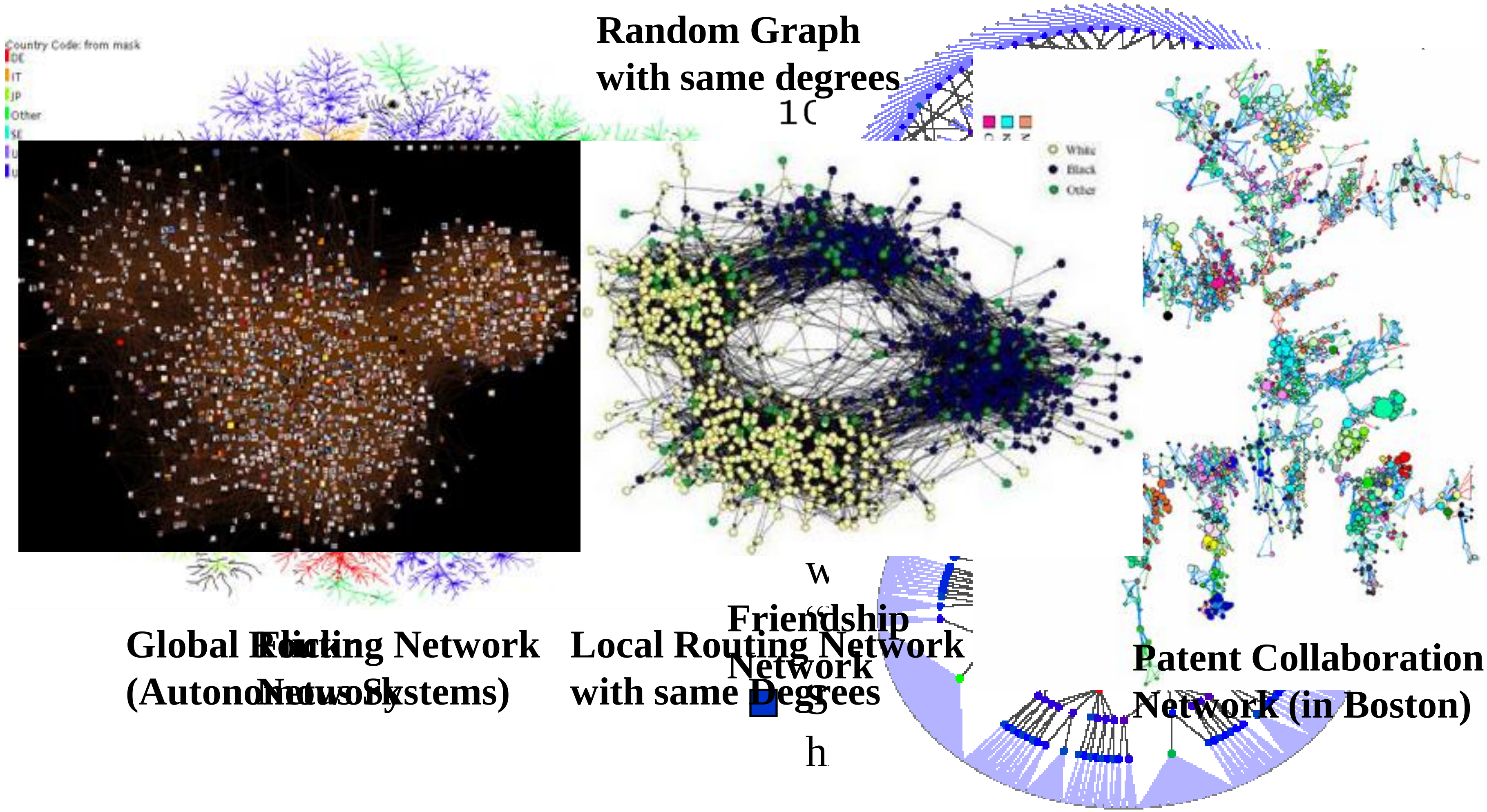


- Sparse graphs
  with large degree-variance.
  "Power-law" degree distributions.
- Small-world, i.e. small diameter,
  high clustering coefficients.

scaling **Web N.0**

The Internet is constantly growing and evolving giving rise to new models and algorithmic questions.

However, in practice, there are discrepancies …



A rich theory of **power-law random graphs** has been developed [**Evolutionary, Configurational Models, & e.g. see Rick Durrett's '07 book].**

"Flexible" models for complex networks:

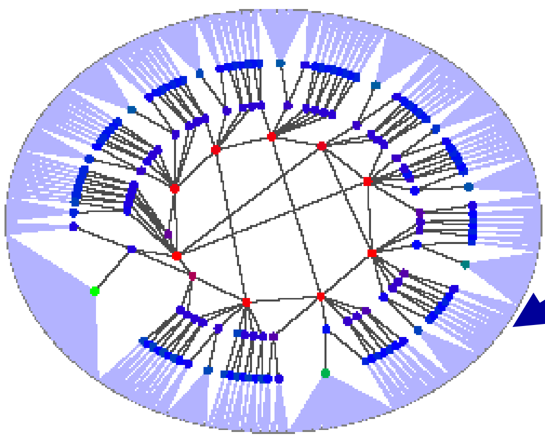
exhibit a "large" increase in the properties of generated graphs

by introducing a "small" extension in the parameters of the generating model.

Case 1: Structural/Syntactic Flexible Model

- Models with power law and arbitrary degree sequences with additional constraints, such as specified joint degree distributions (from random graphs, to graphs with very low entropy).

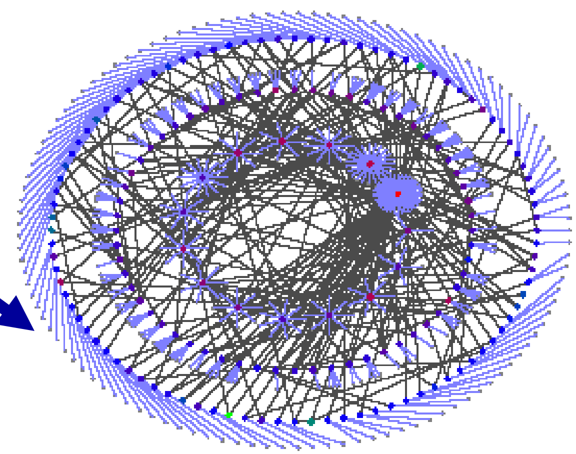
Modifications and Generalizations of Erdos-Gallai / Havel-Hakimi



Assortativity:

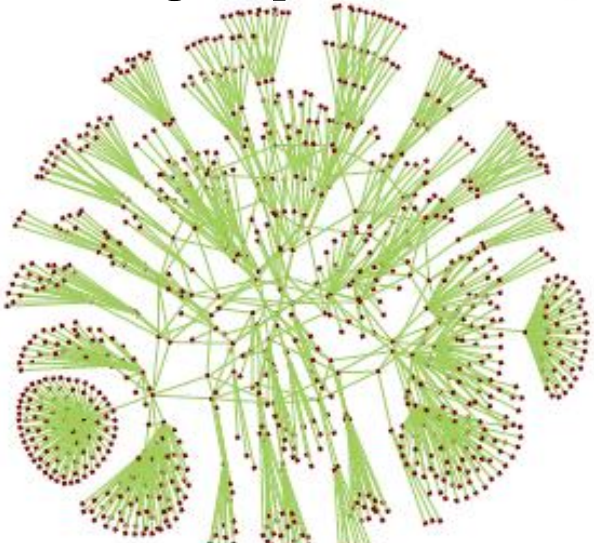
$$\Sigma_{(i,j)\in E} d_i d_j$$

small **large**

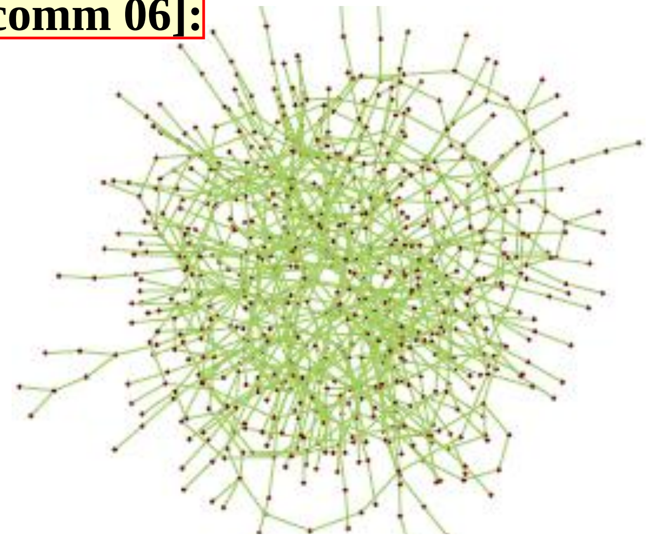


The networking community proposed that **[Sigcomm 04, CCR 06 and Sigcomm 06]**, beyond the degree sequence $d_1 \geq d_2 \geq \ldots \geq d_n$, models for networks of routers should capture how many nodes of degree d_i are connected to nodes of degree d_j.

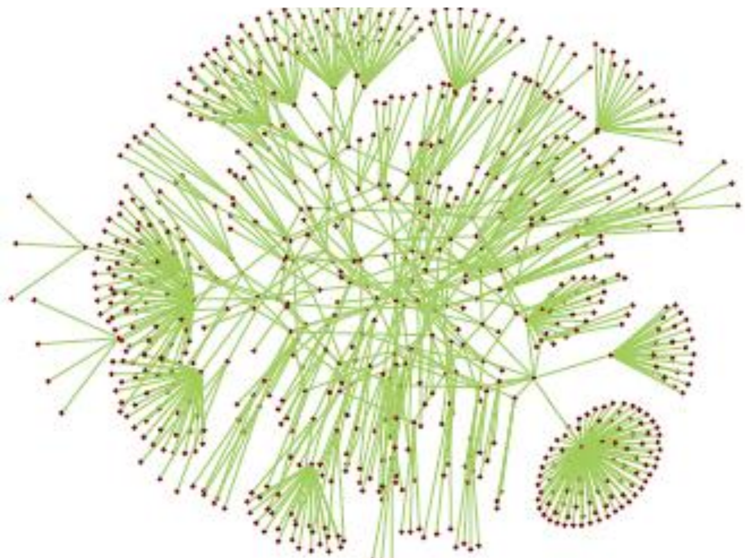
Networking Proposition [CCR 06, Sigcomm 06]:



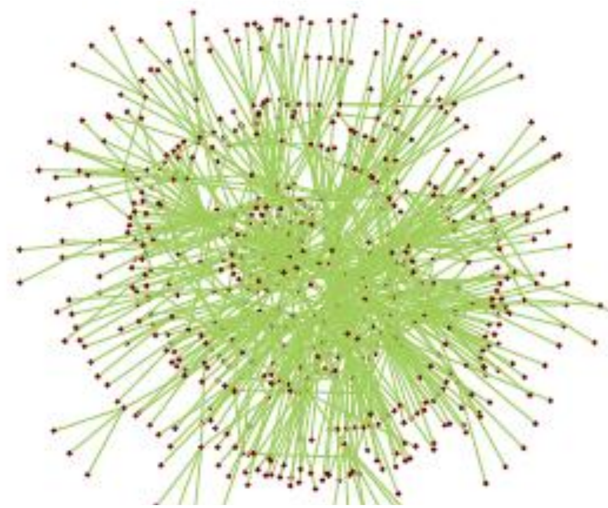
A real highly optimized network G.



A random graph with same average degree as G.



A graph with same number of links between nodes of degree d_i and d_j as G.



A random graph with same degree sequence as G.

The Joint-Degree Matrix Realization Problem is:

Given $< \mathbf{V}, \mathbf{d}, D >$, is there/cnstrct simple graph:
all vertices in V_i have degree $\mathbf{d}(V_i)$, and
there are d_{ij} edges between V_i and V_j
(resp. d_{ii} edges inside V_i).

connected, mincost, random

Definitions

Let $V = [n]$.
Let $\mathbf{V} = \{V_1, \ldots, V_k\}$ denote a partition of V to classes of vertices of the same degree.
Let $\mathbf{d} : \mathbf{V} \to \mathbf{N}$ denote the degrees of each class V_i.
Let $D = (d_{ij})$ be a $k \times k$ matrix, where
d_{ij} is the number of edges between V_i and V_j,
and d_{ii} is the number of edges entirely in V_i.

The (well studied) Degree Sequence Realization Problem is:

Let $V = [n]$. Let $d_1 \geq d_2 \geq \ldots \geq d_n$.
Is there/construct a simple graph on n vertices
with degrees: $d_1 \geq d_2 \geq \ldots \geq d_n$.

connected, mincost, random

The Joint-Degree Matrix Realization Problem is:

Given $< \mathbf{V}, \mathbf{d}, D >$, is there a simple graph where:
all vertices in V_i have degree $\mathbf{d}(V_i)$, and
there are d_{ij} edges between V_i and V_j
(resp. d_{ii} edges inside V_i), $1 \leq i, j \leq k$.

connected, mincost, random

Theorem [Amanatidis, Green, M '08]:

The natural necessary conditions for an instance $< \mathbf{V}, \mathbf{d}, D >$ to have a realization are also sufficient (and have a short description).

The natural necessary conditions for an instance $< \mathbf{V}, \mathbf{d}, D >$ to have a connected realization are also sufficient (no known short description).

There are polynomial time algorithms to construct a realization and a connected realization of $< \mathbf{V}, \mathbf{d}, D >$, or produce a certificate that such a realization does not exist.

Advantages: Flexibility in:

Degree Sequence Realization Problem:

enforcing or precluding certain edges,

Given: arbitrary $d_1 \geq d_2 \geq \ldots \geq d_n$

adding costs on edges and finding mincost realizations.

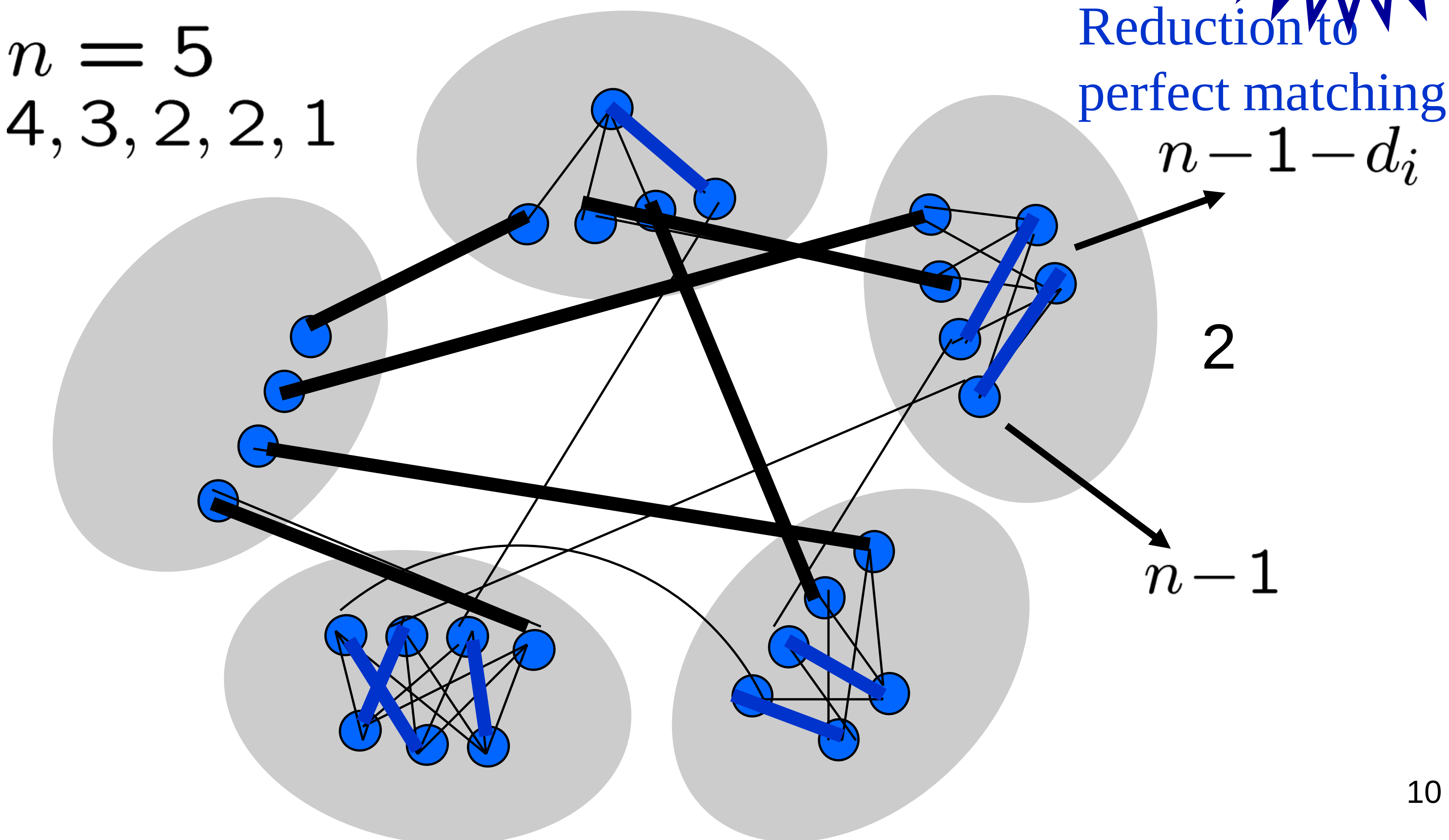
Is this degree sequence realizable ?

close to matching → close to sampling/random generation.

If so, construct a realization.

connected, mincost, random

Reduction to perfect matching:



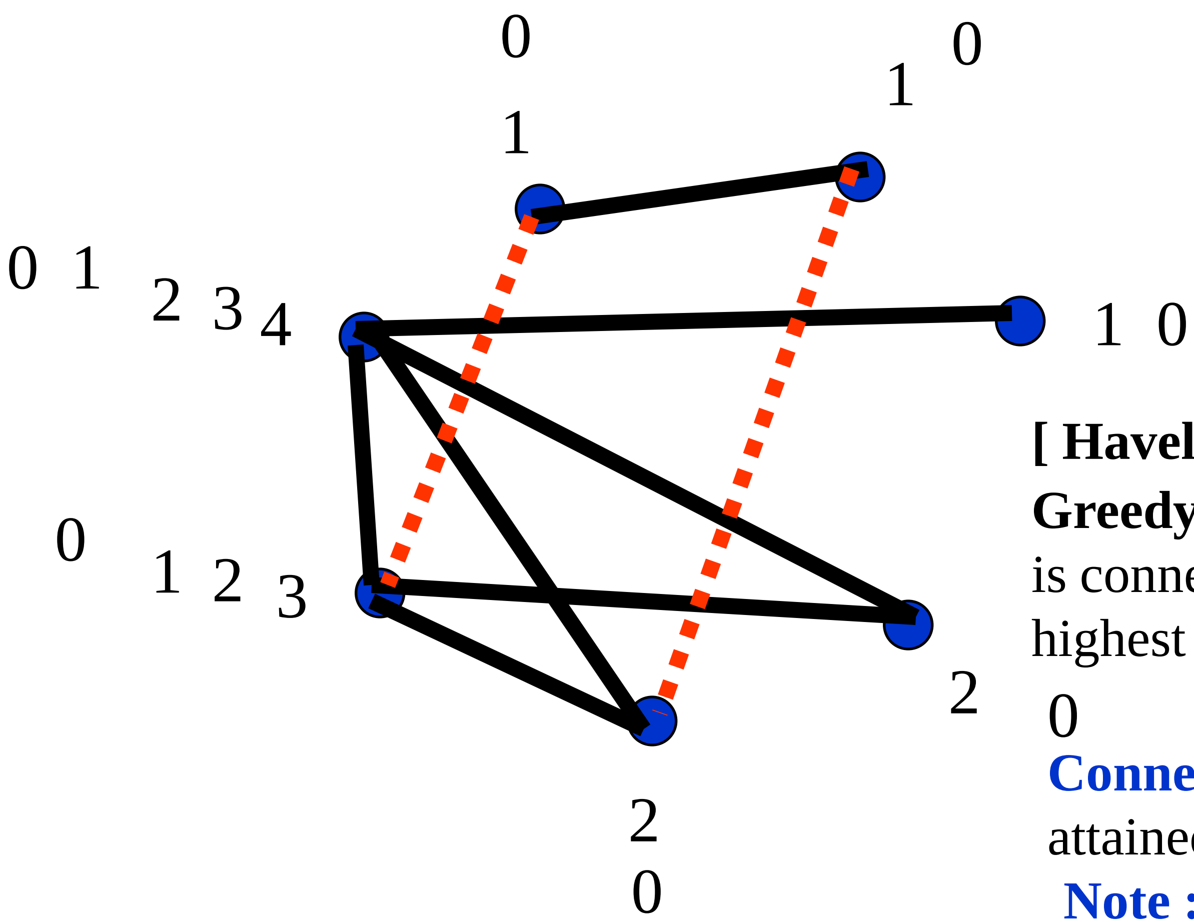
Theorem [**Erdos-Gallai**]:

A degree sequence $d_1 \geq d_2 \geq \ldots \geq d_n$ is realizable

iff the natural necessary condition holds: $\Sigma_{i=1}^{k} d_i \leq k(k-1) + \Sigma_{i=k+1}^{n} \min\{k, d_i\}$

moreover, there is a connected realization $\Sigma_{i=1}^{n} d_i \geq 2\,(n-1)$

iff the natural necessary condition holds:



[Havel-Hakimi] Construction:

Greedy: any unsatisfied vertex is connected with the vertices of highest remaining degree requirements.

Connectivity, if possible, attained with **2-switches** .

Note :all 2-switches are legal .

Theorem, Joint Degree Matrix Realization [Amanatidis, Green, M '08]:

Let $V = [n]$. Then $<\mathbf{V}, \mathbf{d}, D>$ has a graphic realization if and only if:

(i) *Degree Feasibility* holds :

$$2d_{ii} + \Sigma_{j \in [k], j \neq i} d_{ij} = |V_i| \cdot \mathbf{d}(V_i),\ \forall 1 \leq i \leq k.$$

(ii) *Matrix Feasibility* holds: D is symmetric with nonnegative integral entries,

and $d_{ij} \leq |V_i| \cdot |V_j|,\ \forall 1 \leq i < j \leq k$,

while $d_{ii} \leq |V_i| \cdot (|V_i| - 1)/2,\ \forall 1 \leq i \leq k$.

Moreover, when $<\mathbf{V}, \mathbf{d}, D>$ is realizable, there is a polynomial (in n) time algorithm that produces a graphic realization of $<\mathbf{V}, \mathbf{d}, D>$.

Proof [sketch]:

Necessity is obvious.

Sufficiency follows from the **greedy** polynomial time construction algorithm outlined next.

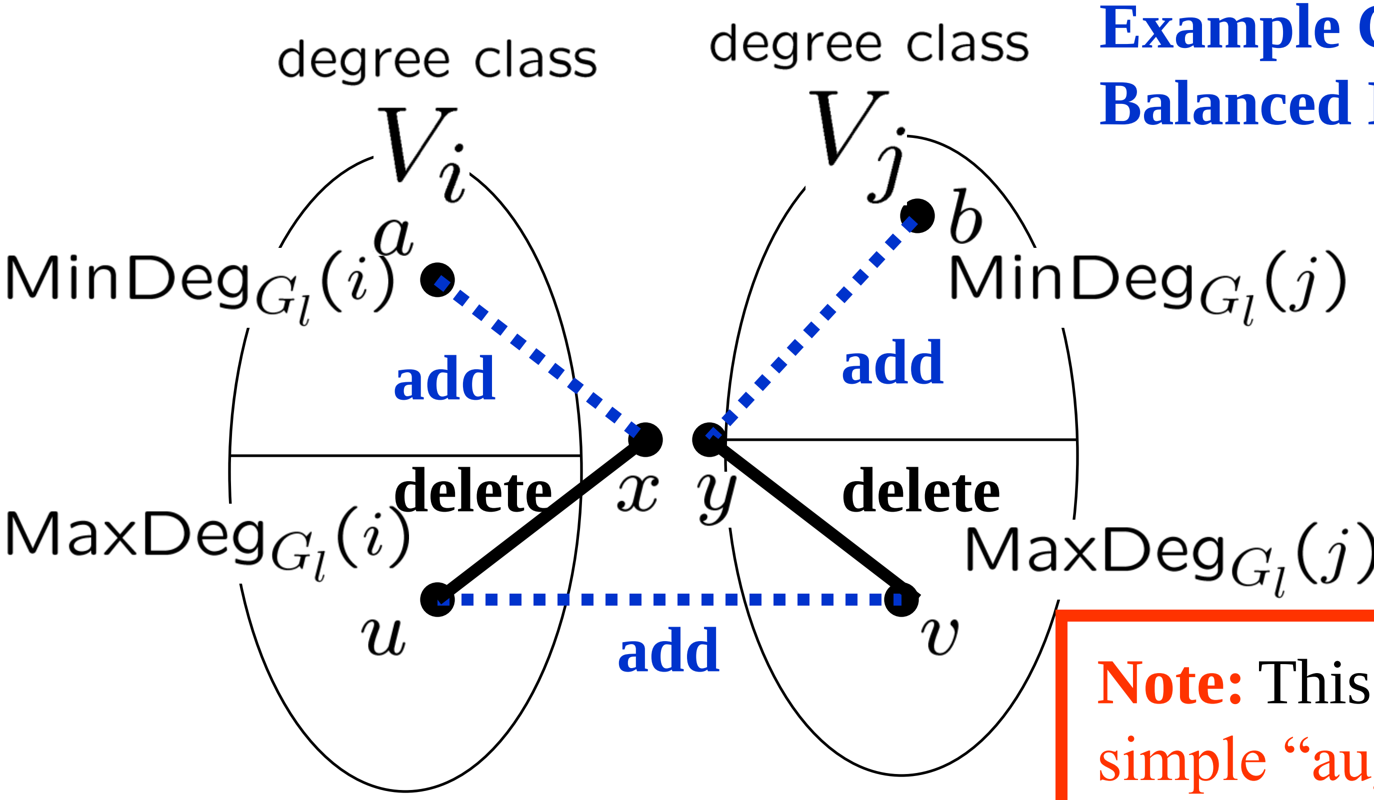
Balanced Degree Invariant:

The key idea of the algorithm is to maintain balanced degrees within each degree class.

In particular, where G_l is the graph after the l-th iteration, the algorithm maintains:

$$\max_{v \in V_i} \deg_{G_l}(v) - \min_{v \in V_i} \deg_{G_l}(v) \leq 1\,,\ \forall 1 \leq i \leq k\,.$$

Example Case Maintaining Balanced Degree Invariant:



Note: This may NOT be a simple "augmenting" path.

Theorem, Joint Degree Matrix Connected Realization [Amanatidis, Green, M '08]:

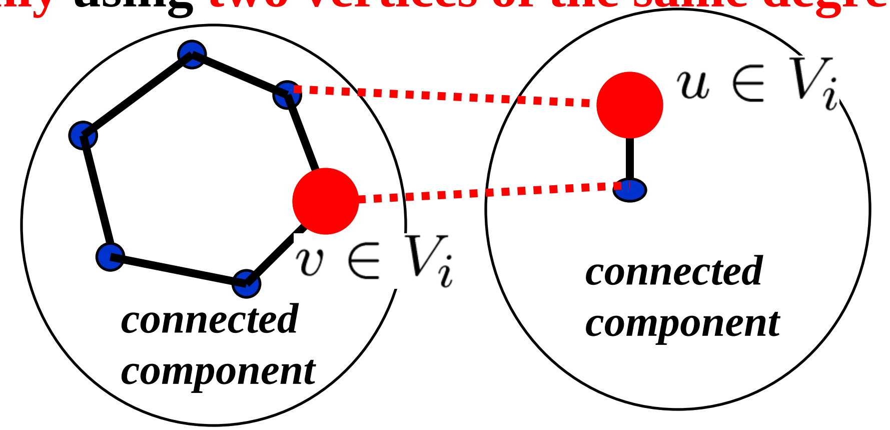
Let $V = [n]$. Let $< \mathbf{V}, \mathbf{d}, D >$ be a realizable instance of the degree matrix realization problem. Then, there is a polynomial (in n) time algorith that, either outputs a *connected* graphic realization of $< \mathbf{V}, \mathbf{d}, D >$, or outputs a *certificate* that a connected graphic realization of $< \mathbf{V}, \mathbf{d}, D >$ does not exist.

Proof [remarks]:

We do not know of a polynomially short description of necessary and sufficient conditions.

The algorithm explores vertices of the same degree in different components, in a recursive manner.

Main Difficulty: Two connected components are amenable to rewiring by 2-switches, only using two vertices of the same degree.



Open Problems for Joint Degree Matrix Realization

- Construct mincost and/or random realization,
  or connected realization.
- Satisfy constraints between arbitrary subsets of vertices.
- Is there a reduction to matchings or flow or
  some other well understood combinatorial problem?
- Is there evidence of hardness ?
- Is there a simple generative model for graphs
  with low assortativity ? (explanatory or other …)

Talk Outline

Complex Networks in Web N.0

Flexible (further parametrized) Models

1. Structural/Syntactic Flexible Models

2. Semantic Flexible Models

Models & Algorithms Connection : Kleinberg's Model(s) for Navigation

Distributed Searching Algorithms with Additional Local Info/Dynamics

1. On the Power of Local Replication

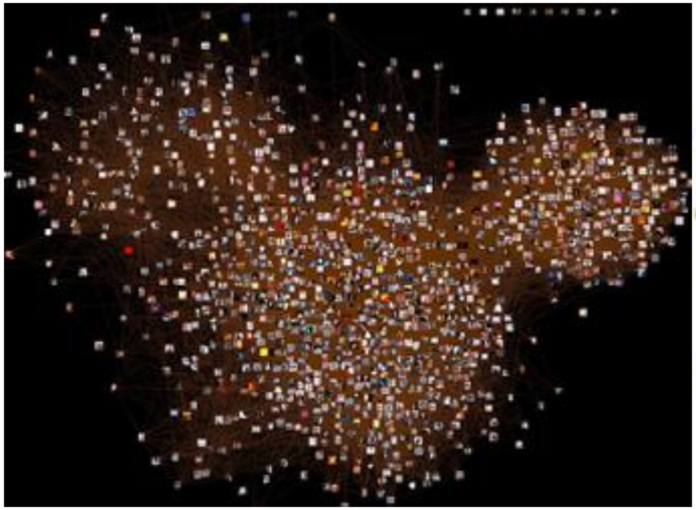
2. On the Power of Topology Awareness via Link Criticality

Conclusion : Web N.0 Model & Algorithm characteristics:
further parametrization, typically local,
locality of info in algorithms & dynamics.
Dynamics become especially important.

ODD

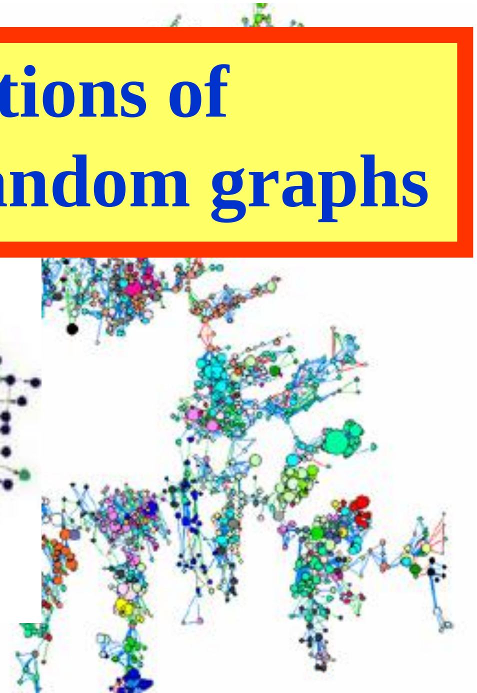
Case 2: Semantic Flexible Model(s)

Generalizations of Erdos-Renyi random graphs



Flickr Network

Friendship Network



Patent Collaboration Network (in Boston)

- Models with semantics on nodes, and links among nodes with semantic proximity generated by very general probability distributions.

Varying structural characteristics

- **RANDOM DOT PRODUCT GRAPHS**
- **KRONECKER GRAPHS**

Also densification, shrinking diameter, … C. Faloutsos, MMDS08

RANDOM DOT PRODUCT GRAPHS

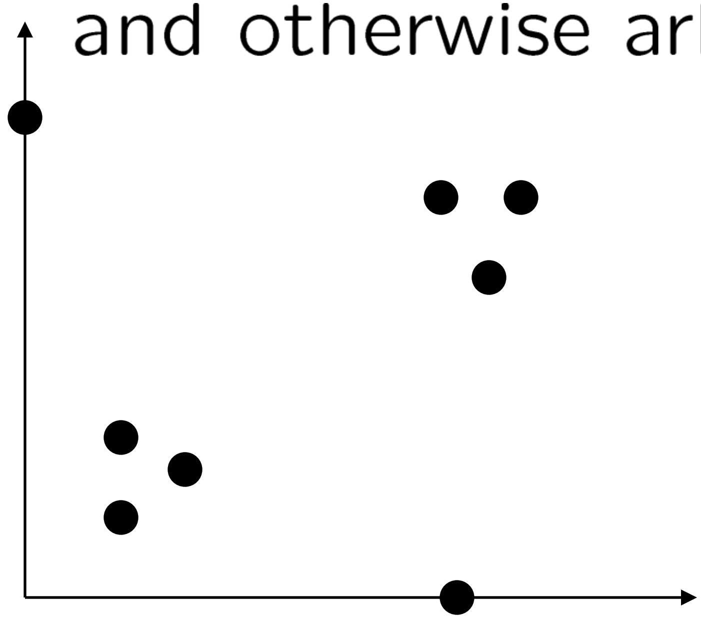
Kratzl,Nickel,Scheinerman 05
Young,Scheinerman 07
Young,M 08

The Model $G_g^{\langle\cdot,\cdot\rangle}(\mathbf{X}, n)$

n vertices each generated according to $\mathbf{X}$

each vertex is a vector in d-dim space
one coordinate for each attribute
d is fixed

$\mathbf{X}$ is a probability distribution in $\mathbf{R}^d$
$\mathbf{X}$ is in the positive orthant $[0, 1/\sqrt{d}]$
and otherwise arbitrary



$d = 2$
$n =$ very large
scales

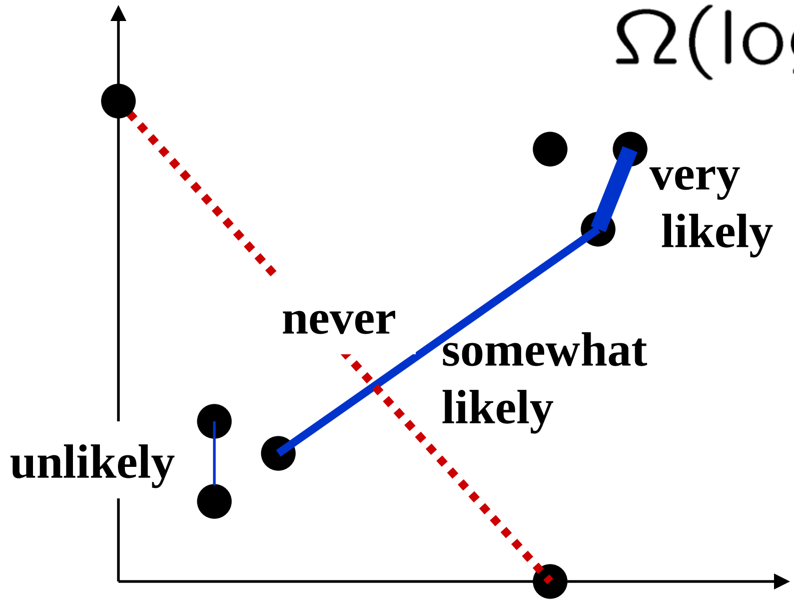
The Model $G_g^{\langle\cdot,\cdot\rangle}(\mathbf{X}, n)$

edges are added between two vertices
with probabilities proportional
to their inner product,denoting similarity,

and inversely proportional
to a non-decreasing "sparsification" function

$$g = g(n)$$

$$\Omega(\log n) \leq g(n) \leq O(n)$$



$d = 2$

$n =$ very large scales

SUMMARY OF RESULTS

- A semi-closed formula for degree distribution

Model can generate graphs with a wide variety of densities

average degrees $\Omega(\log n)$ up to $O(n)$.

and wide varieties of degree distributions, including power-laws.

- Diameter characterization :

Determined by Erdos-Renyi for similar average density,

if all coordinates of **X** are in $(0, 1/\sqrt{d})$ (will say more about this).

- Positive clustering coefficient,

depending on the “distance” of the generating distribution from the uniform distribution.

Remark: Power-laws and the small world phenomenon are the hallmark of complex networks.

A Semi-closed Formula for Degree Distribution

Let $\omega \in [0,1]$ be a random variable distributed as $\left\langle \frac{\mathbb{E}[\mathbf{X}]}{\|\mathbb{E}[\mathbf{X}]\|}, \mathbf{X} \right\rangle$

Theorem [Young, M '08] For any valid $\mathbf{X}$ on $\mathbb{R}^d$, $d \geq 1$, let v be a vertex in $G = G_g^{\langle\cdot,\cdot\rangle}(\mathbf{X}, n)$, Let $0 < \delta, \epsilon < 1$ be such that $(1+\delta)(1-\epsilon) > 1$. Then,

$$\mathbb{P}(|\deg(v) - k| \leq \delta k) \leq \min\left\{ ((1-\epsilon)e^{\epsilon})^{(1-\delta)k} + ((1+\epsilon)e^{-\epsilon})^{(1+\delta)k}, \frac{2(1+\delta^2)n}{(g(n)\epsilon(1-\delta^2)k)^2} \right\} + \int_{(1-\epsilon)(1-\delta)t_n^k}^{(1+\epsilon)(1+\delta)t_n^k} d\omega$$

$$\mathbb{P}(|\deg(v) - k| \leq \delta k) \geq \left(1 - \min\left\{ (2(1+\epsilon)e^{-\epsilon})^{(1-\delta)k}, \frac{2n}{(g(n)\epsilon(1-\delta)k)^2} \right\}\right) \int_{(1+\epsilon)(1-\delta)t_n^k}^{(1-\epsilon)(1+\delta)t_n^k} d\omega$$

$$t_n^k = \frac{g(n)k}{\|\mathbb{E}[\mathbf{X}]\|(n-1)}$$

Theorem (removing error terms) [Young, M '08]

$$\mathbf{P}(|\deg(v) - k| \leq \delta k) \simeq \int_{(1-\delta)\frac{g(n)k}{\|\mathbf{E}[\mathbf{X}]\|n}}^{(1+\delta)\frac{g(n)k}{\|\mathbf{E}[\mathbf{X}]\|n}} d\omega$$

Example:

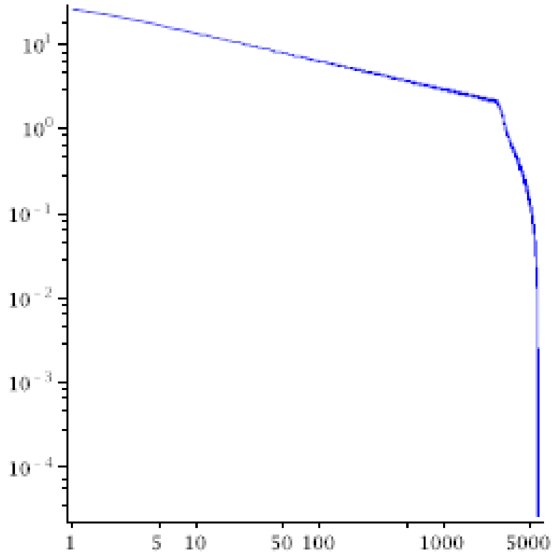
Consider the one dimensional random dot product graph with distribution $\mathbf{Pr}(x \leq r) \leq r^{1/\alpha}$ $\alpha \geq 1$

and various densification functions.

for $\sqrt{\frac{n}{g(n)}} \leq k \leq \frac{(1+\alpha)(n-1)}{g(n)(1+\delta)}$ (a wide range of degrees, except for very large degrees)

$$\mathbb{P}(|\deg(v) - k| \geq \delta k) \geq c_{n,\alpha} \left((1+\delta)^{\frac{1}{\alpha}} - (1-\delta)^{\frac{1}{\alpha}} \right) k^{\frac{1}{\alpha}-1}.$$

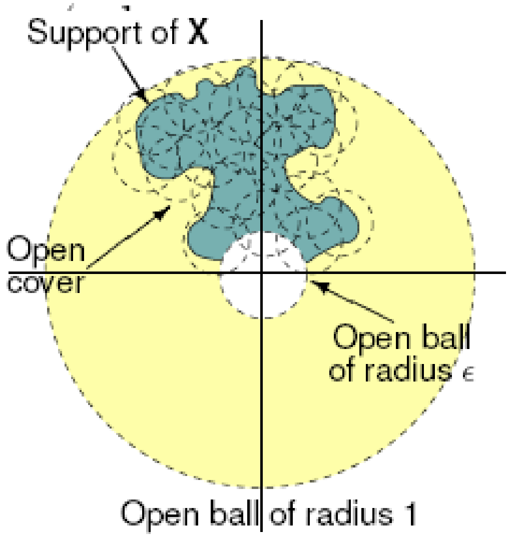
$$\mathbf{Pr}(\deg(v) \simeq k) \simeq k^{\frac{1}{\alpha}-2}$$

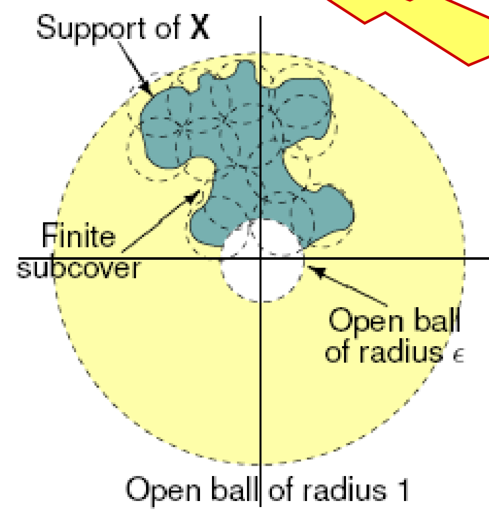


This is in agreement with real data.

Diameter Characterization

We have obtained a method of lifting results about the diameter of the Erdős-Rényí model to $G_g^{\langle\cdot,\cdot\rangle}(\mathbf{X}, n)$. Specifically, using the boundedness of the support of $\mathbf{X}$, we can prove that if Erdős-Rényí model $\mathcal{G}\left(\Theta\left(\frac{1}{g(n)}\right), n\right)$ has low diameter, then the diameter of $G_g^{\langle\cdot,\cdot\rangle}(\mathbf{X}, n)$ is not much bigger. For this result only, we assume that $\mathbf{X} \in (0, 1/\sqrt{d})$





d is fixed
$\mathbf{X}$ is fixed
is critical
in proof

Remark: If $\mathbf{X} \in [0, 1/\sqrt{d}]$ the graph can become disconnected
It is important to obtain characterizations of connectivity as $\mathbf{X}$ approaches $[0, 1/\sqrt{d}]$ .This would enhance model flexibility

Clustering Characterization

Theorem [Young, M '08] For vertices, u, v, and w in $G_g^{\langle\cdot,\cdot\rangle}(\mathbf{X}, n)$, $\mathbb{P}(u \sim w \mid u \sim v, v \sim w) \geq \mathbb{P}(u \sim w)$, with equality holding if and only if $\mu_{\mathbf{X}}(\mathbb{E}[\mathbf{X}]) = 1$, that is $\mathbf{X}$ is almost surely constant.

Remarks on the proof

Clustering depends on
the distance of $G_g^{\langle\cdot,\cdot\rangle}(\mathbf{X}, n)$ from a standard Erdős-Rényi model.

Clustering depends on "size" of cov($\mathbf{X}$).

$$\mathrm{cov}(\mathbf{X}) = \mathbb{E}\left[\mathbf{X}\mathbf{X}^T\right] - \mathbb{E}[\mathbf{X}]\,\mathbb{E}[\mathbf{X}]^T$$

is symmetric positive semidefinite

may assume coordinates have covariance 0.

Open Problems for Random Dot Product Graphs

- Fit real data, and isolate "benchmark" distributions $\mathbf{X}$.
- Characterize connectivity (diameter and conductance) as $\mathbf{X}$ approaches $[0, 1/\sqrt{d}]$.
- Do/which further properties of $\mathbf{X}$ characterize further properties of $G_g^{\langle\cdot,\cdot\rangle}(\mathbf{X}, n)$?
- Evolution: **X** as a function of *n* ? (including: two connected vertices with small similarity, either disconnect, or increase their similarity).
- Should/can $d = \log n$?
- Similarity functions beyond inner product (e.g. Kernel functions).
- Algorithms: navigability, information/virus propagation, etc.

KRONECKER GRAPHS [Faloutsos, Kleinberg,Leskovec 06]

0	**1**
1	**1**

1-bit vertex character-ization

0	**0**	**0**	**1**
0	**0**	**1**	**1**
0	**1**	**0**	**1**
1	**1**	**1**	**1**

2-bit vertex character-ization

0	**0**	**0**	**0**	**0**	**0**	**0**	**1**
0	**0**	**0**	**0**	**0**	**0**	**1**	**1**
0	**0**	**0**	**0**	**0**	**1**	**0**	**1**
0	**0**	**0**	**0**	**1**	**1**	**1**	**1**
0	**0**	**0**	**1**	**0**	**0**	**0**	**1**
0	**0**	**1**	**1**	**0**	**0**	**1**	**1**
0	**1**	**0**	**1**	**0**	**1**	**0**	**1**
1	**1**	**1**	**1**	**1**	**1**	**1**	**1**

3-bit vertex character-ization

log n-bit vertex characterization

Another "semantic" " flexible" model, introducing parametrization.

STOCHASTIC KRONECKER GRAPHS

[Faloutsos, Kleinberg, Leskovec 06]

a	b
b	c

aa	ab	ba	bb
ab	ac	bb	bc
ba	bb	ca	cb
bb	bc	cd	cc

aaa	aab	aba	abb	baa	bab	bba	bbb
aab	aac	abb	abc	bab	bac	bbb	bbc
aba	abb	aca	acb	bba	bbb	bca	bcb
abb	abc	acd	acc	bbb	bbc	bcd	bcc
baa	bab	bba	bbb	caa	cab	cba	cbb
bab	bac	bbb	bbc	cab	cac	cbb	cbc
bba	bbb	bca	bcb	cba	cbb	cca	ccb
bbb	bbc	bcd	bcc	cbb	cbc	ccd	ccc

$$0 \leq \mathbf{a}, \mathbf{b}, \mathbf{c} \leq 1$$

Several properties characterized (e.g. multinomial degree distributions, densification, shrinking diameter, self-similarity).

Large scale data set have been fit efficiently !

Talk Outline

Complex Networks in Web N.0

Flexible (further parametrized) Models

1. Structural/Syntactic Flexible Models

2. Semantic Flexible Models

Models & Algorithms Connection : Kleinberg's Model(s) for Navigation

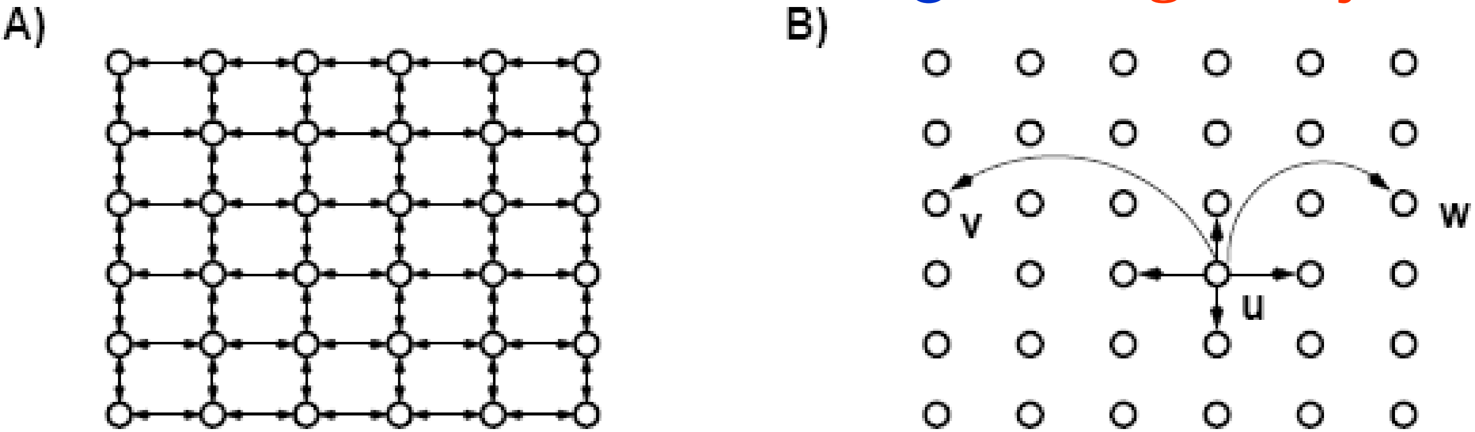
Distributed Searching Algorithms with Additional Local Info/Dynamics

1. On the Power of Local Replication

2. On the Power of Topology Awareness via Link Criticality

Conclusion : Web N.0 Model & Algorithm characteristics:
Further Parametrization, Locality of Info & Dynamics.

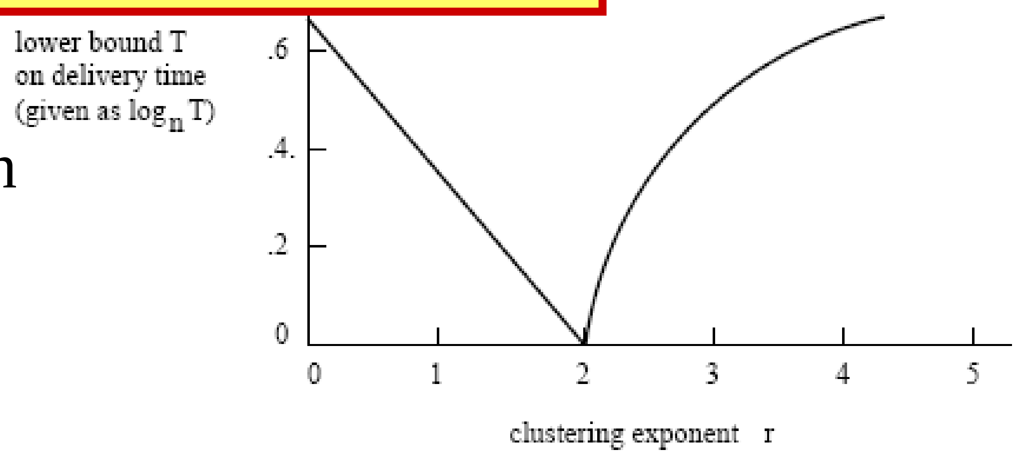
Where it all started: Kleinberg's navigability model



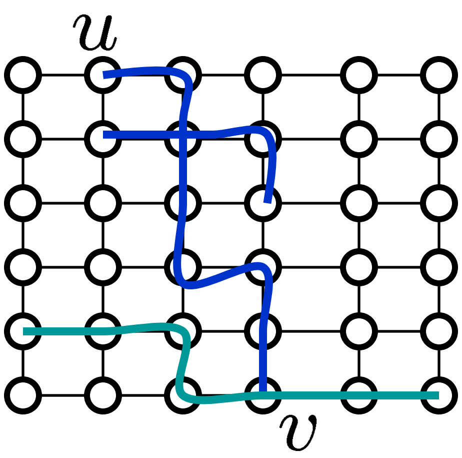
$$\mathrm{Pr}[\{u,v\}] \simeq \mathrm{dist}(u,v)^{-r}$$

Moral: Parametrization is essential in the study of complex networks

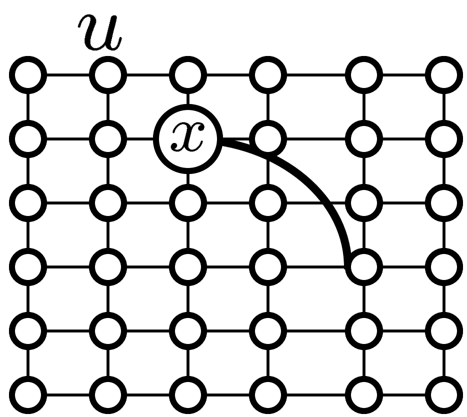
Theorem [Kleinberg]: The only value for which the network is navigable is ***r* =2**.



Strategic Network Formation Process **[Sandberg 05]:**



all pairs of vertices u and v
choose a random u-v shortest path



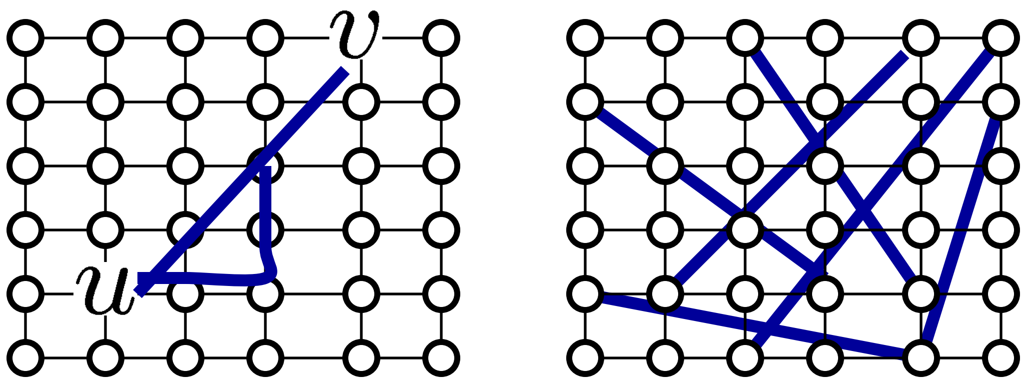
each node x computes :
for each node $u \neq x$
$P(u)=$paths through x
with endpoint u

each node x adds link to node u
with probability $\simeq P(u)$

Experimentally, the resulting network has structure and navigability similar to Kleinberg's small world network.

Strategic Network Formation Process [Green & M '08]

simplification of **[Clauset & Moore 03]:**



repeat
simeoultaneously

- each node u is presented a uniformly random node u
- u starts navigating to v
- if the navigation steps exceed L then u adds a link to v

until no links are added

Experimentally, the resulting network becomes navigable after $\text{poly}\log n$ steps but does not have structure similar to Kleinberg's small world network.

Talk Outline

Complex Networks in Web N.0

Flexible (further parametrized) Models

1. Structural/Syntactic Flexible Models

2. Semantic Flexible Models

Models & Algorithms Connection : Kleinberg's Model(s) for Navigation

Distributed Searching Algorithms with Additional Local Info/Dynamics

1. On the Power of Local Replication

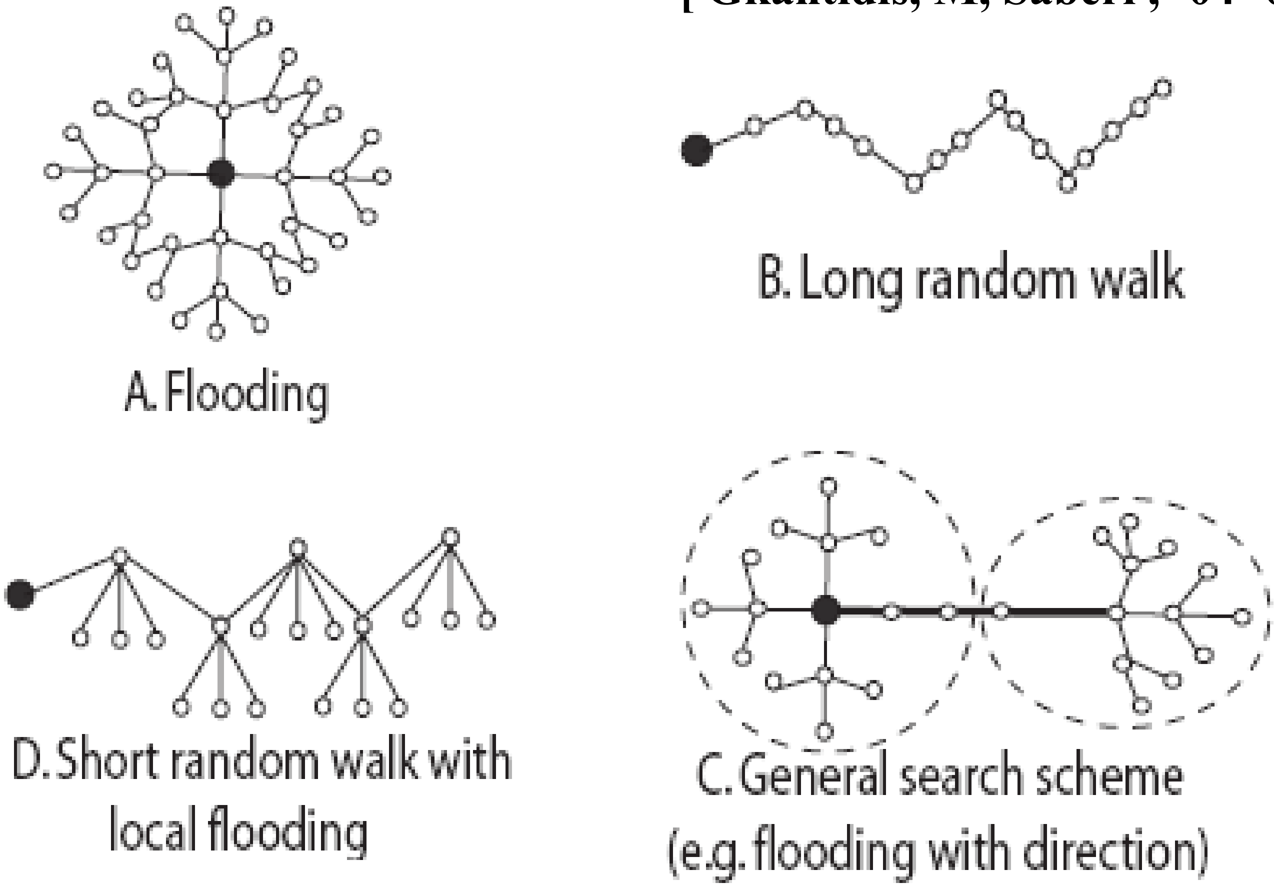
2. On the Power of Topology Awareness via Link Criticality

Conclusion : Web N.0 Model & Algorithm characteristics:
further parametrization, typically local,
locality of info in algorithms & dynamics.
Dynamics become especially important.

1. On the Power of Local Replication

How do networks **search** (propagate information) :

[Gkantidis, M, Saberi , ‘04 ‘05]

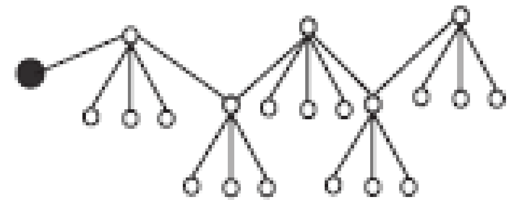


Cost = queried nodes / found information

1. On the Power of Local Replication

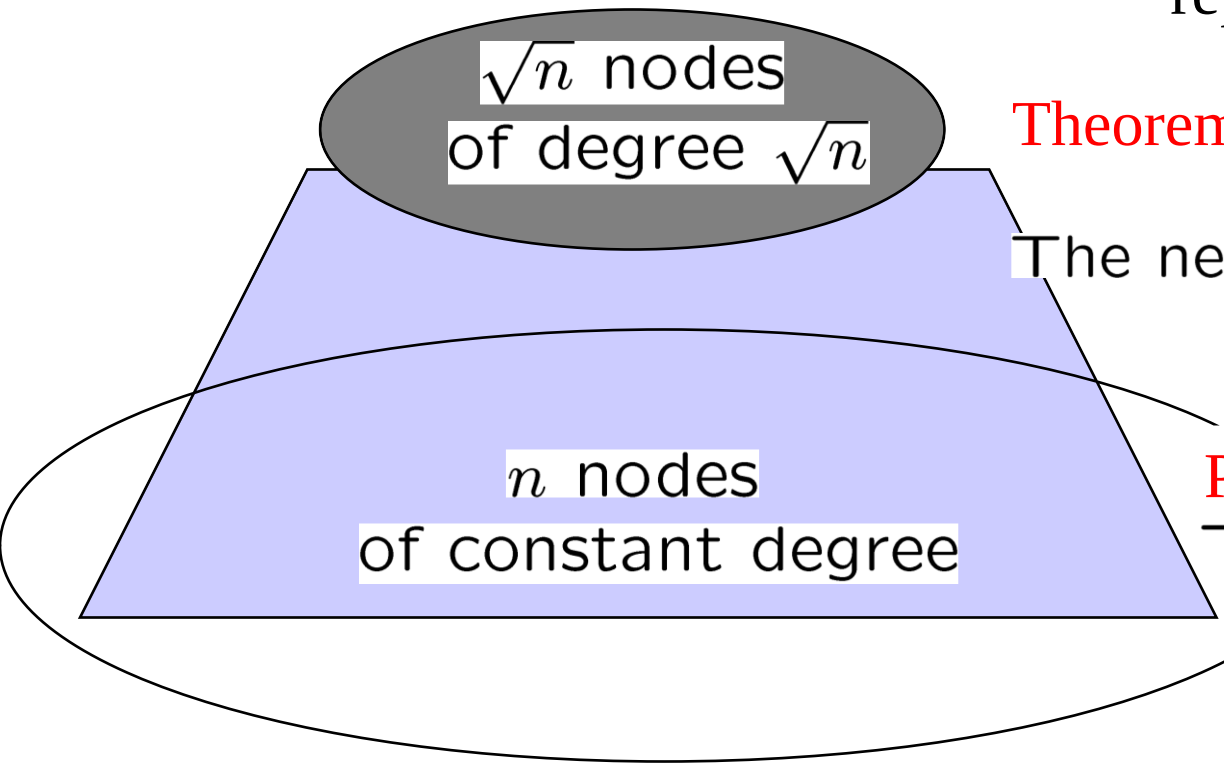
[Gkantidis, M, Saberi , '04 '05]

[M, Saberi , Tetali '05]



D.Short random walk with local flooding

network=random graph



Equivalent to one-step local replication of information.

Theorem (extends to power-law random graphs):

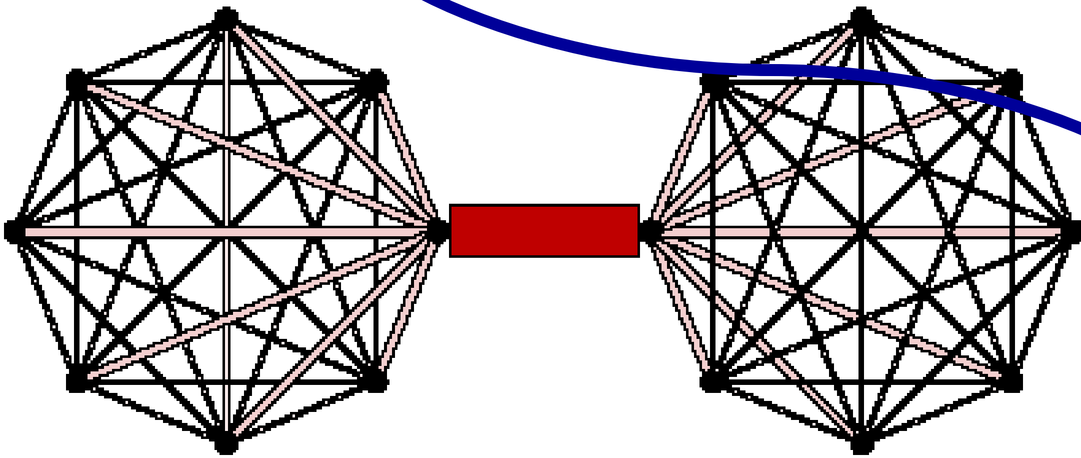
The network can be searched by RANDOM WALK in $\tilde{O}(\sqrt{n})$ steps.

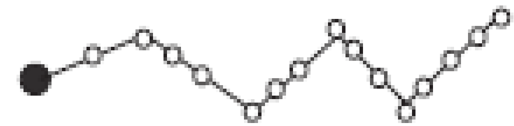
Proof :

The cover time of a random graph on $\sqrt{n}$ nodes.

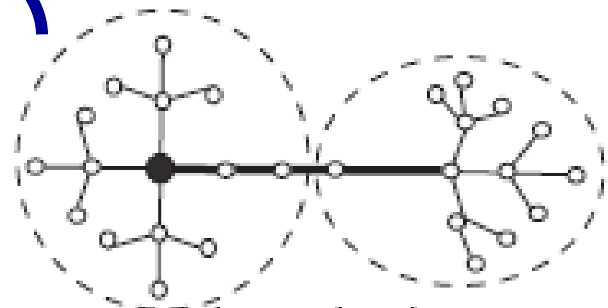
notice: $O(n)$ nodes, $O(n)$ links

2. On the Power of Topology Awareness via **Link Criticality**

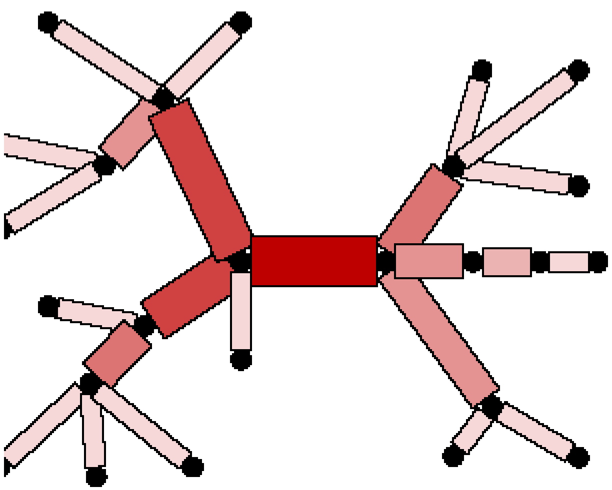


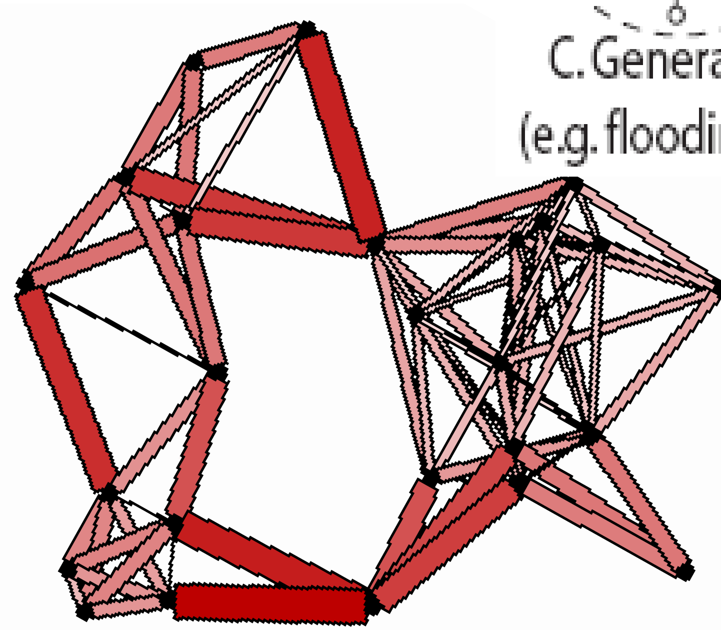


B. Long random walk

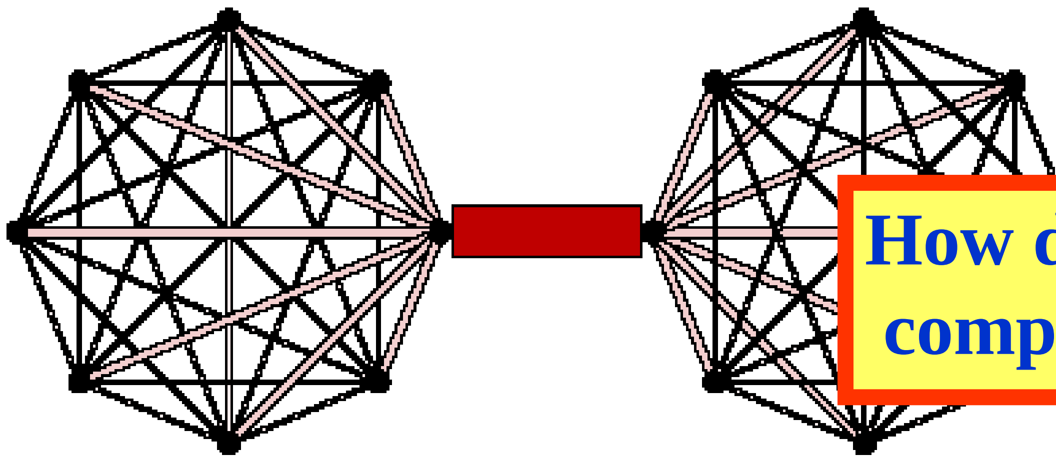


C. General search scheme
(e.g. flooding with direction)





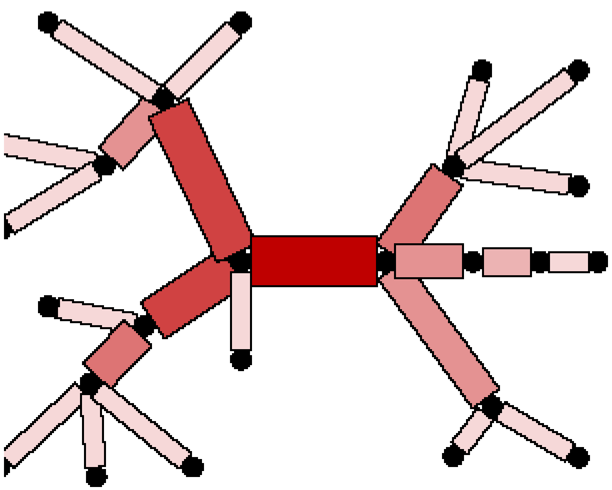
2. On the Power of Topology Awareness via **Link Criticality**

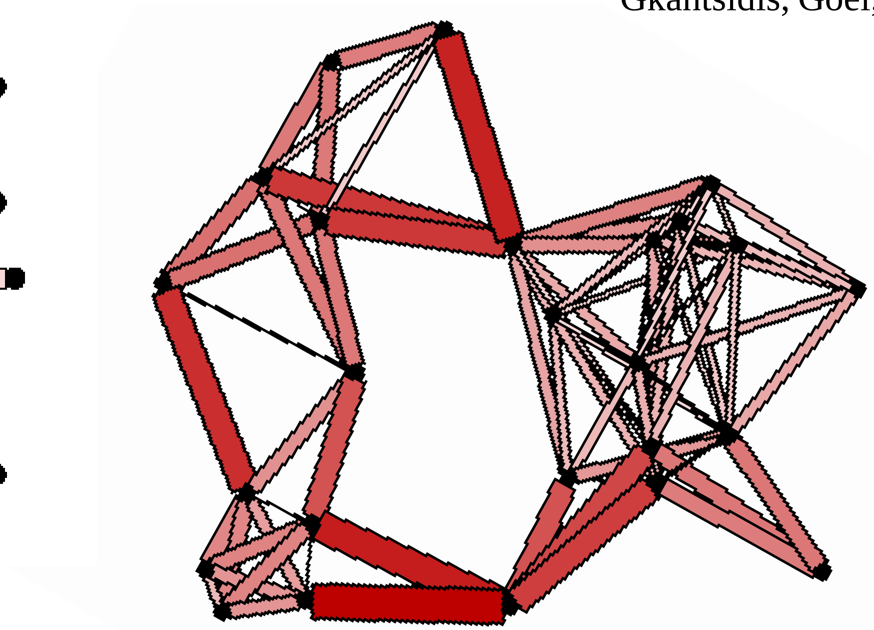


Via Semidefinite Programming
"Fastest Mixing Markov Chain"

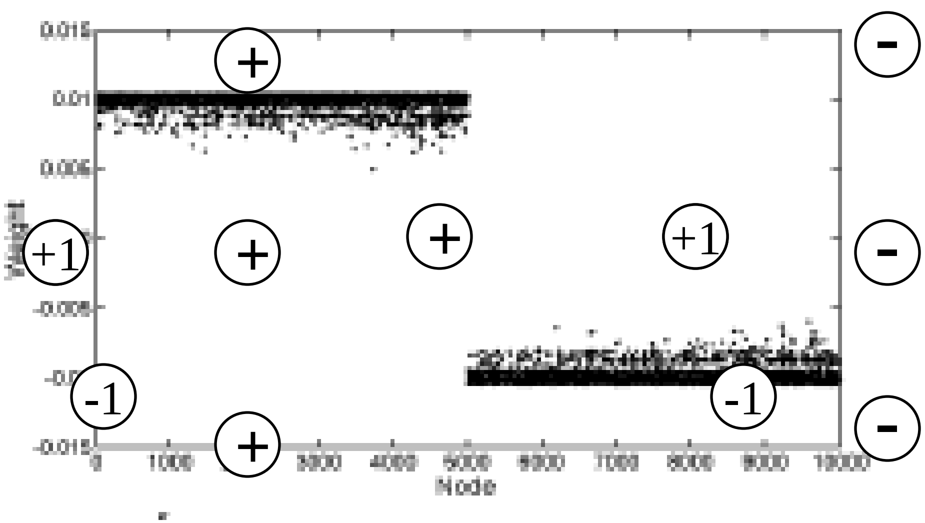
How do social networks compute link criticality ?

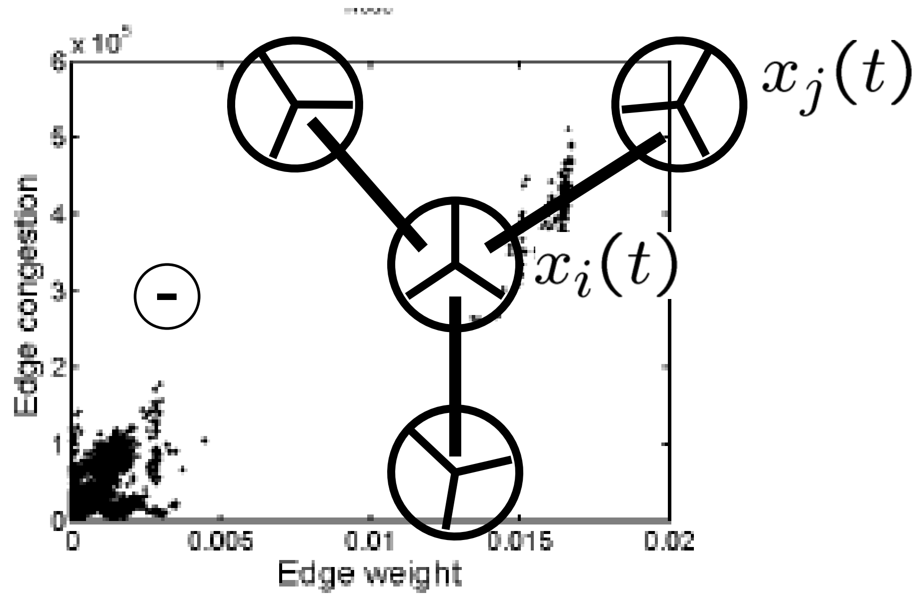
of Principal Eigenvector(s),
Distributed, Asynchronous
Gkantsidis, Goel, Mihail, Saberi 07





Link Criticality via Distributed Asynchronous Computation of Principal Eigenvector(s) [Gkantsidis,Goel,M,Saberi 07]





Start with $(x_1, \ldots, x_n) \perp (1, \ldots, 1)$

Step: For all nodes asynchronously $x_i(t+1) = \sum_{(i,j)\in E} \frac{x_i(t)+x_j(t)}{2d}$

For links, $+ \frac{x_j(t)}{2d}$

Hardest part: Numerical Stability.

Talk Outline

Complex Networks in Web N.0

Flexible (further parametrized) Models

1. Structural/Syntactic Flexible Models

2. Semantic Flexible Models

Models & Algorithms Connection : Kleinberg's Model(s) for Navigation

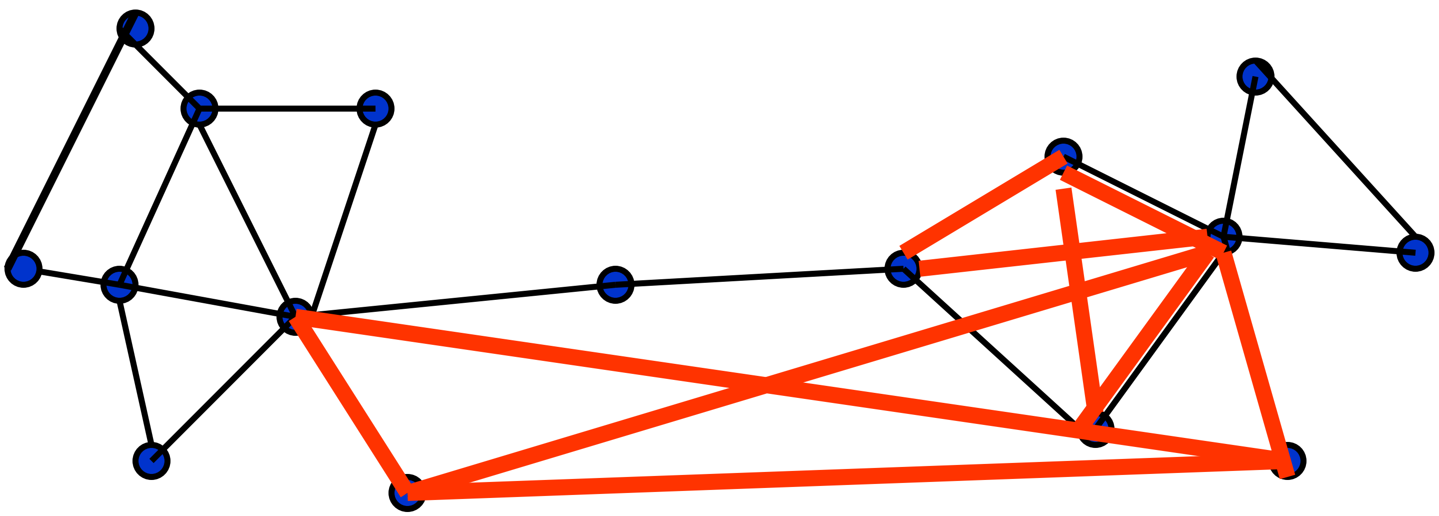
Distributed Searching Algorithms with Additional Local Info/Dynamics

1. On the Power of Local Replication

2. On the Power of Topology Awareness via Link Criticality

Conclusion : Web N.0 Model & Algorithm characteristics:
further parametrization, typically local,
locality of info in algorithms & dynamics.
Dynamics become especially important.

Topology Maintenance = Connectivity & Good Conductance



Theorem [**Feder,Guetz,M,Saberi 06**]:The Markov chain corresponding to a **local 2-link switch** is rapidly mixing if the degree sequence enforces diameter at least 3, and for some $d \leq n/2$, $\frac{d+1}{n-d} d \leq d_i \leq d$.