Structural sparsity of complex networks

Felix Reidl, Peter Rossmanith, Fernando Sánchez Villaamil, Blair D. Sullivan* and Somnath Sikdar

Theoretical Computer Science

RWTH AACHEN

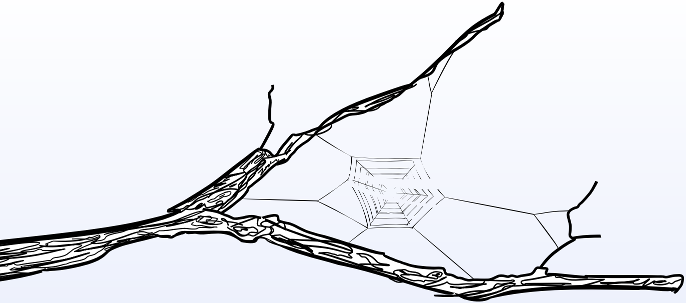
*North Carolina State University

@Finse 2014

Contents

- Costa, Rodrigues, Travieso, Villas Boas, Characterization of Complex Networks: A survey of measurements. 2008
- Newman, The structure and function of complex networks. 2003
- Albert & Barabási, Statistical mechanics of complex networks. 2002
- Dorogovtsev & Mendes, Evolution of networks. 2001

Complex Networks



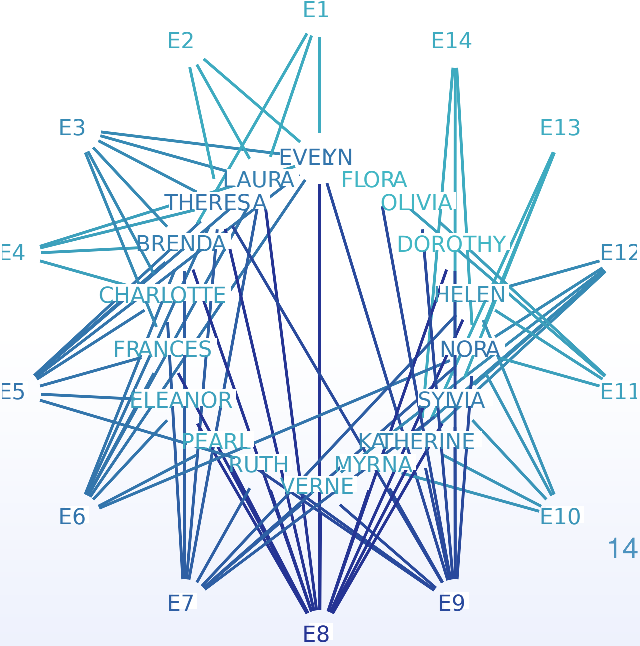
A certainly incomplete history

1734 Euler: Königsberger Brücken
1920 First mapping of social networks by social scientists
1950 Simon: ‘Rich get richer’
1959 Erdős & Rényi: On random graphs
1965 Price: Citation network is scale-free
1967 Milgram: Six degrees of separation
1994 Wassermann & Faust: Clustering coefficient
(under different name)
1995 Molloy & Reed: Rigorious notion of degree sequences
1998 Watts & Strogatz: Comparative study of networks
1999 Barabási & Albert: Rediscover and improve Price’s work
2000 Kleinberg: Small-world routing

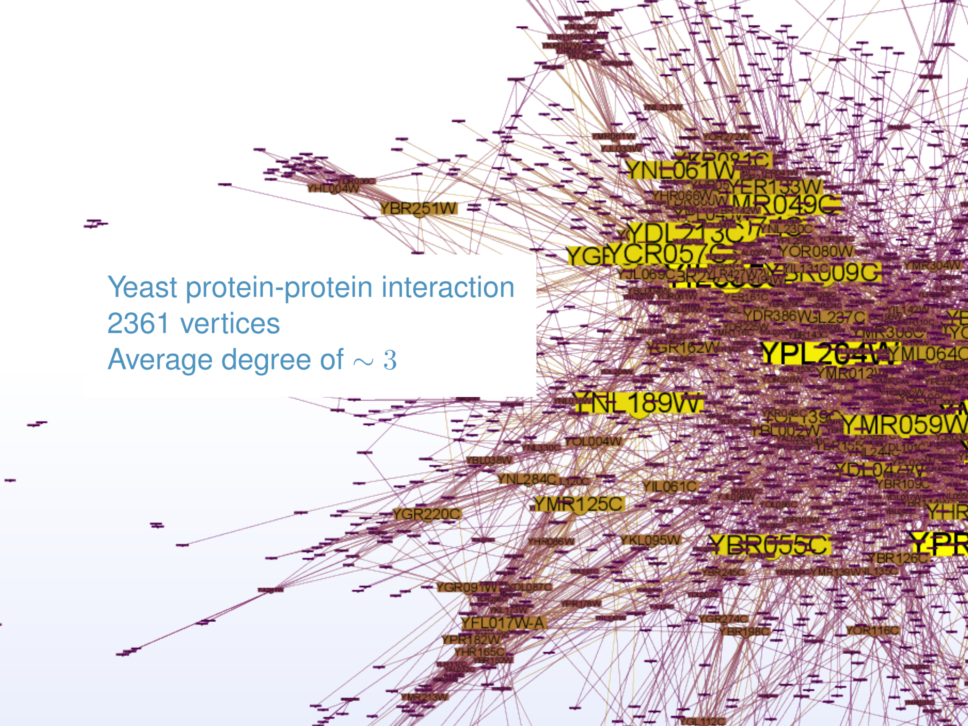
Networks are graphs as they appear in the “real world”

A big field

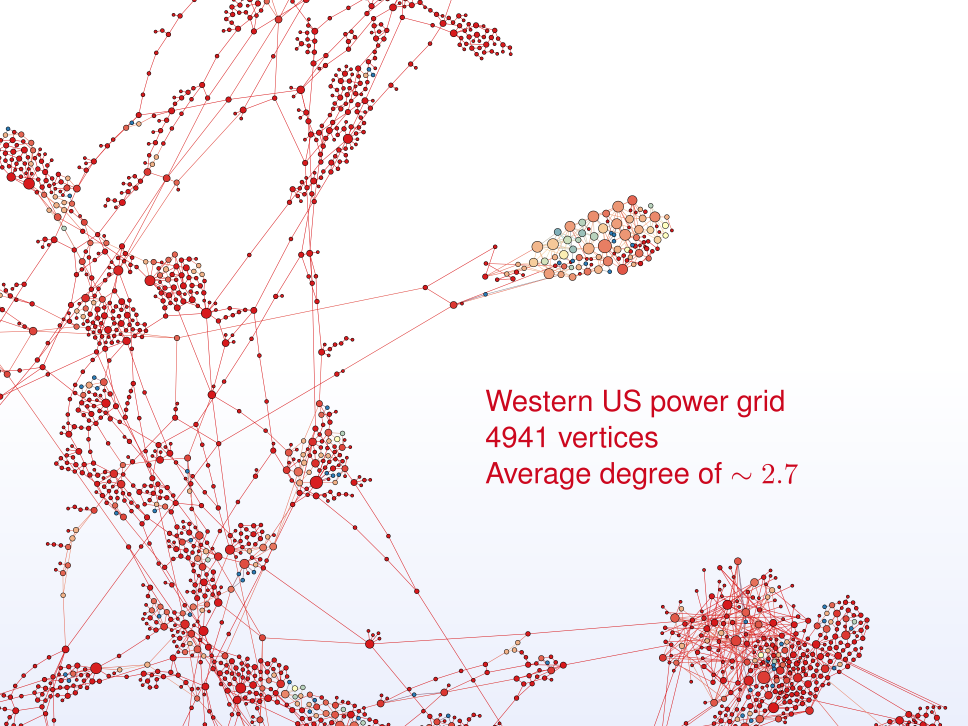
Social	Biology
Friendship	Food webs
Co-authorship	Neural networks
Sexual contacts	Protein-protein interaction
Movie actors	Cell metabolism
Telephone calls	Protein folding states
Infrastructure	**Other**
Power grid	Word co-occurence
Internet	Software packages
Railway networks	Synonyms
Electric circuits	Spacetime...?



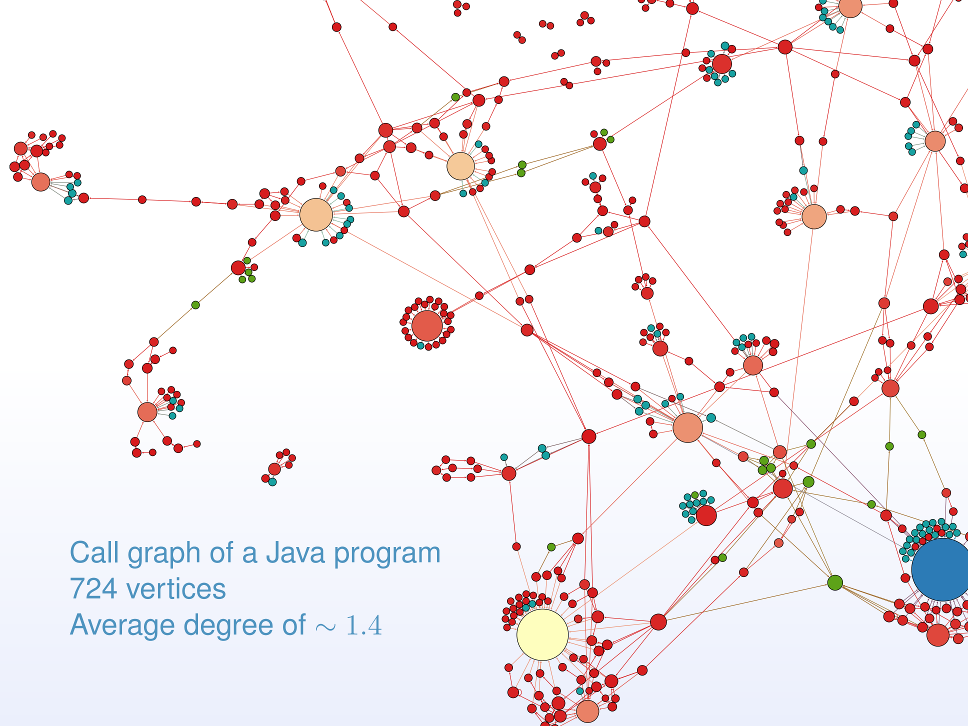
Southern Women
Davis et al., 1930
18 women
14 events over 9 month



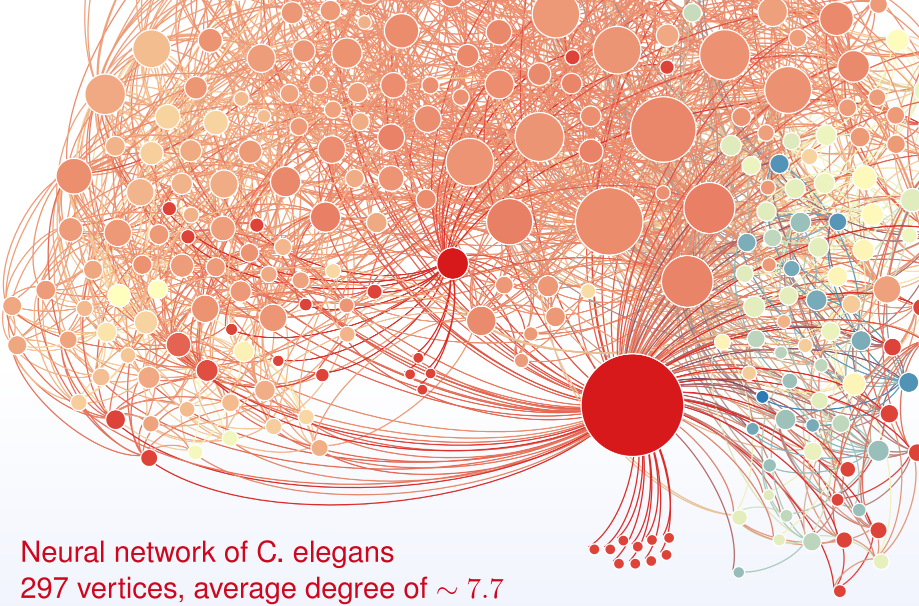
Yeast protein-protein interaction
2361 vertices
Average degree of ~ 3



Western US power grid
4941 vertices
Average degree of ~ 2.7



Call graph of a Java program
724 vertices
Average degree of ~ 1.4



Neural network of C. elegans
297 vertices, average degree of ~ 7.7

Central questions about networks

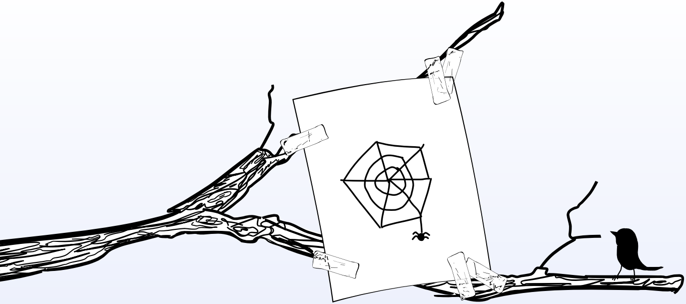
Network topology

- How are vertices connected?
- Diameter, average path length
- Which vertices are ‘important’?
- Navigation or mixing in networks
- Community detection
- Network resilience
- ...

Network recognition
How to distinguish networks or fingerprint them.

Network evolution
How do networks change over time?

Modeling complex networks



Networks models

Models have three goals:

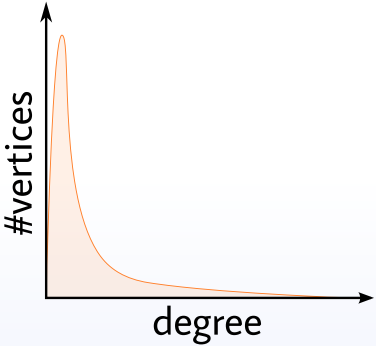
1. Insight into underlying process
2. Handle for mathematical theorems
3. Provide test data

Depending on the emphasis, models are vastly different.

No one-size-fits-all!

Two important observations

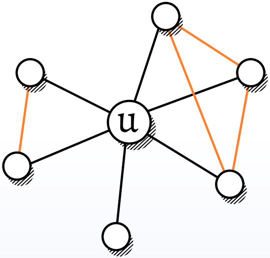
Degree distribution



Power-law for many networks:

$$P(k) \sim k^{-\gamma}$$

Clustering

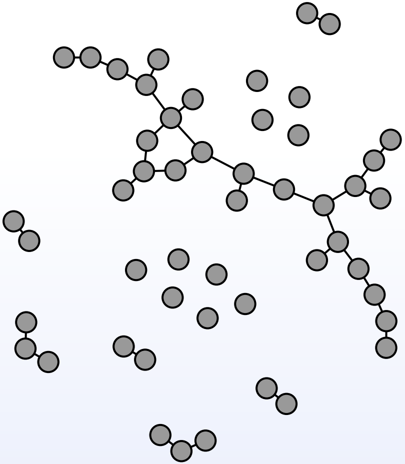


Number of triangles divided by number of triples consistent for similar networks.

Erdős-Rényi

$G(n,p)$: n-vertex graph in wich every edge is present with probability p. For sparse graphs, we want $np = O(1)$.

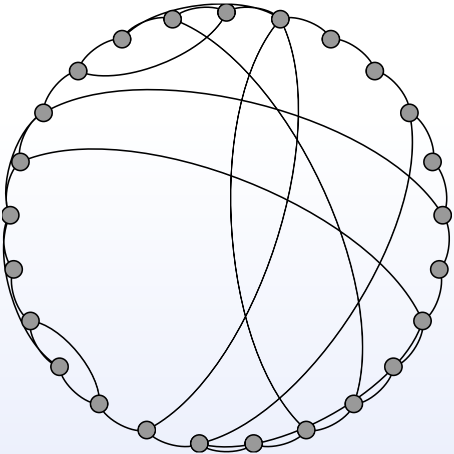
- Well-understood
- Simple model
- Clustering $\sim p$
- Degree distribution too symmetric



Watts-Strogatz

Parameters n, k, p: create a n-vertex cycle where every vertex is connected to the $k/2$ previous and next vertices. Rewire every edges with probability p.

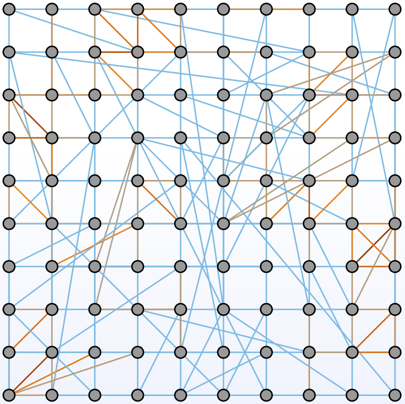
- Small-world
- Clustering independent of size
- Average degree unrealistic
  (usually $k > \log n$)



Kleinberg

Start with a $\sqrt{n} \times \sqrt{n}$ grid-like graph. For every vertex v, add q edges to it, weighing the probability for endpoint w by $\frac{1}{d(u,w)^r}$.

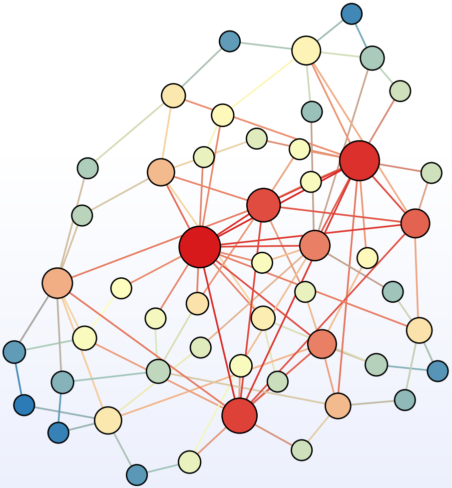
- Small-world *routing*
- Very restrictive
  (designed to model one single aspect)



Barabási-Albert

Rich-get-richer: start with small graph of m_0 vertices. Iteratively add a new vertex, connect it to m old vertices chosen with probabilities proportional to their degree.

- Small-world
- Power-law degree distribution
- Clustering independent of size
- Not very adaptive

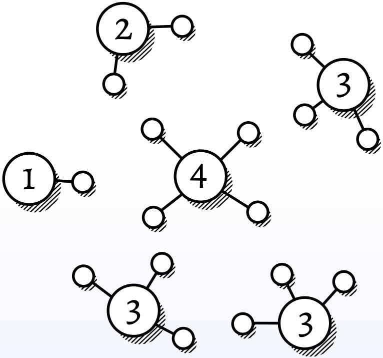


Fixed degree distributions

Instead of trying to achieve a certain degree distribution by designing a model, why not just prescribe it directly?

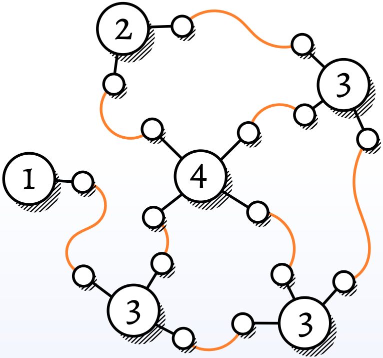
Fixed degree distributions

Instead of trying to achieve a certain degree distribution by designing a model, why not just prescribe it directly?



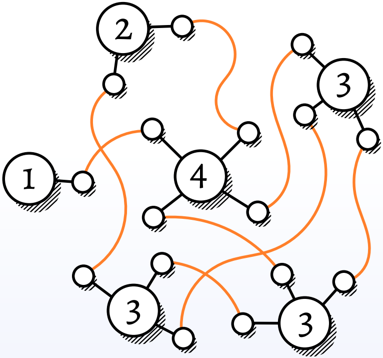
Fixed degree distributions

Instead of trying to achieve a certain degree distribution by designing a model, why not just prescribe it directly?



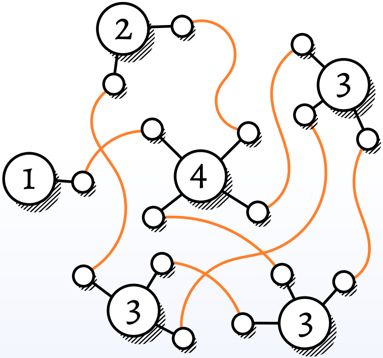
Fixed degree distributions

Instead of trying to achieve a certain degree distribution by designing a model, why not just prescribe it directly?



Fixed degree distributions

Instead of trying to achieve a certain degree distribution by designing a model, why not just prescribe it directly?



How to formalize ‘degree distribution’ rigorously?

Molloy-Reed

Definition

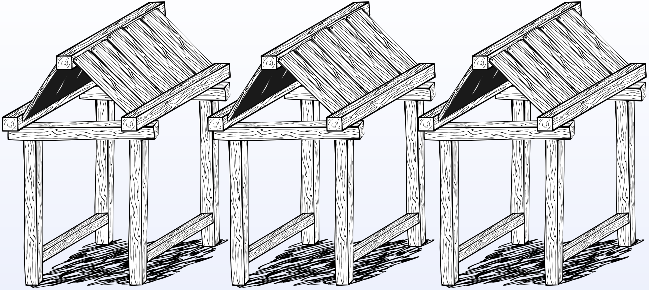
An *asymptotic degree sequence* is a sequence of integer-valued functions $\mathcal{D} = d_0, d_1, d_2, \ldots$ such that for all $n \geq 0$

1. $\sum_{i=0}^{n-1} d_i(n) = n$
2. $d_j(n) = 0$ for $j \geq n$

Molloy-Reed conditions (simplified):

- Feasible: can be realized by a sequence of graphs
- Smooth: $\lim_{n\to\infty} d_i(n)/n = \lambda_i$ for some constant λ_i
- Sparse: $\sum_{i=1}^{\infty} i\lambda_i = \mu$ for some constant μ
- Max-degree: $d_i(n) = 0$ for $i > n^{1/4}$

Structural sparsity



Back to graph theory

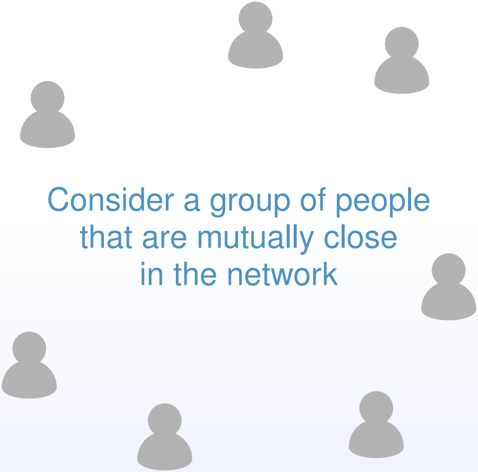
Our fleeting suspicion:
networks are probably sparse in a *structural* sense.
(If they are sparse to begin with)

But in *what* structural sense?

- Low treewidth? Sadly not.
- Planar? Certainly not.
- Bounded-degree? No.
- Exluding a minor/top-minor? Improbable.
- Degenerate? **Very likely!**

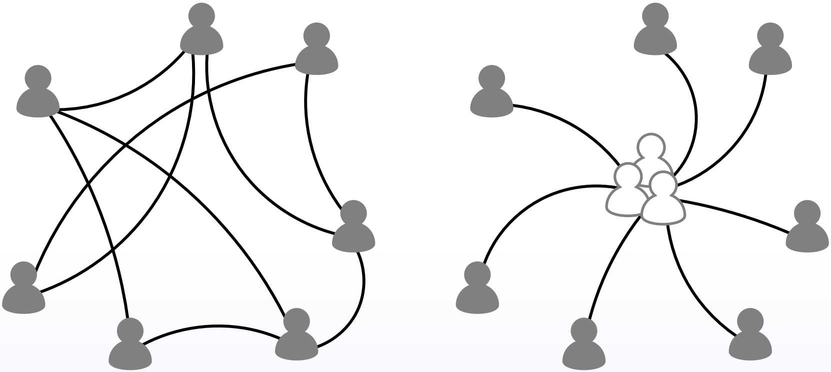
But degenerate graphs have few nice properties. Can we find something a bit more restrictive?

Intuition



Consider a group of people
that are mutually close
in the network

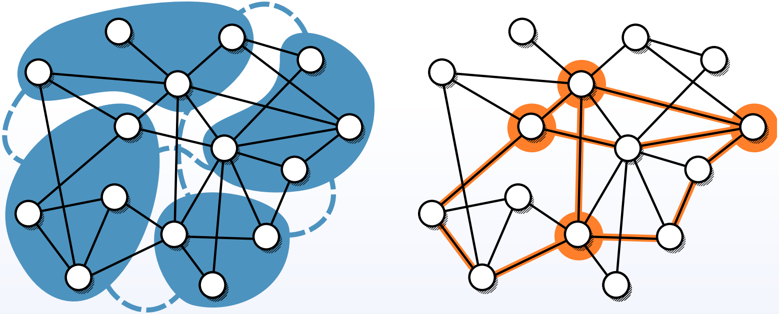
Intuition



Which situation seems more likely?

Bounded expansion

A graph class $\mathcal{G}$ has *bounded expansion* if every r-shallow minor has density at most $f(r)$.

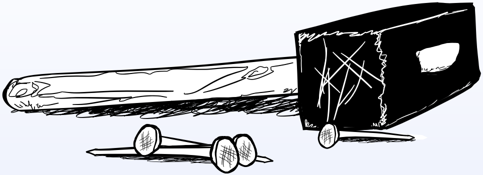


Our (informal) result

1. Graphs created under the Molloy-Reed model have a.a.s. bounded expansion.
2. Adding random edges to a bounded-degree graph with probability bounded by μ/n for some constant μ yields a.a.s. graphs of bounded expansion.

The second result is tight in the sense that adding random edges to a star-forest already gives dense minors with high probability.

Applications



Clustering coefficient

- Idea: number of triangles intrinsic property of network
- Local clustering coefficient of a vertex v:

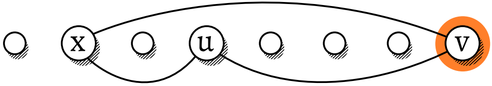
$$c_v = \frac{\#\text{triangles containing } v}{\#P_3\text{s with } v \text{ as center}} = \frac{2|E(N(v)|}{d(v)(d(v)-1)}$$

- Clustering coefficient* of a graph G:

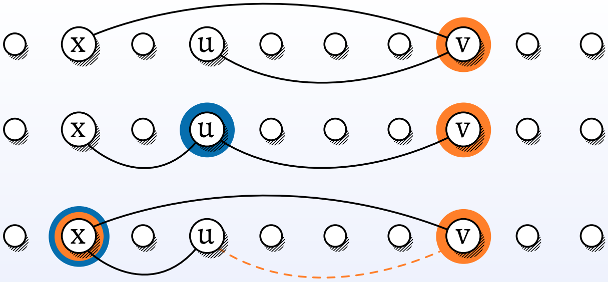
$$C_G = \frac{1}{n} \sum_{v \in V(G)} c_v$$

Counting triangles and P_3s

Degeneracy ordering of vertices: every vertex has at most d neighbours to the left.



Counting triangles: easy. What about P_3s?



Clustering coefficient

- Best known algorithm to count triangles in general: $O(m^{1.41})$ using fast matrix multiplication. (Along, Yuster, Zwick 1997)
- Random sampling, linear-time approximations
- We can do this *with a simple algorithm* in $O(d^2 n)$ time in d-degenerate graphs.
- Similar measures (transitivity) that depend on triangles and P_3s in the same time

Takeaway: if degeneracy is reasonably low, you really want this type of algorithm.

Centrality

- Basic question: how important is a vertex in the network?
- Centrality measure $c\colon V(G) \to \mathbf{R}$
  - Degree-centrality
  - Page-rank
  - Betweeness-centrality
  - Closeness-centrality

Closeness: $c(v) = \sum_{v \neq w \in G} \frac{1}{d(v,w)}$

- Bad: needs all-pairs-shortest paths
- But: Constants-length paths can be handled well in bounded expansion graphs

Truncated closeness: $c_d(v) = \sum_{w \in N^d(v)} \frac{1}{d(v,w)}$

Truncated closeness

Theorem (Nešetřil, Ossana de Mendez)

Let G be a graph of bounded expansion. For every d one can compute in linear time a directed supergraph $\vec{G}_d$ with bounded in-degree and an arc labeling $\omega : \vec{E}(\vec{G}_d) \to \mathbf{N}$ such that for every vertex pair $u, v \in G$ with $d(u, v) \leq d$ one of the following holds:

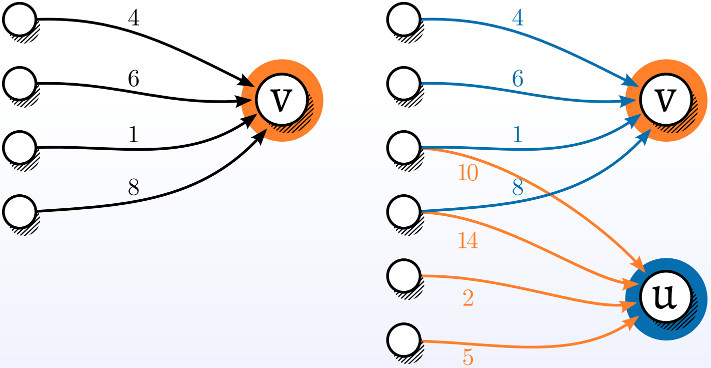
- $uv \in \vec{G}_d$ *and* $\omega(uv) = d(u, v)$
- $vu \in \vec{G}_d$ *and* $\omega(vu) = d(u, v)$
- *there exists* $w \in N^-_{\vec{G}_d}(u) \cap N^-_{\vec{G}_d}(v)$ *such that* $\omega(wu) + \omega(wv) = d(u, v)$

In short: we have a data structure to query short distances in constant time

Truncated closeness

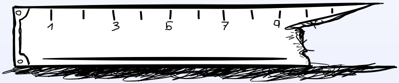
For d-truncated closeness we work on $\vec{G}_d$ in two phases

1. Aggregate distances of direct neighbours in $\vec{G}_d$
2. Aggregate distances of indirect neighbours in $\vec{G}_d$



Truncated closeness

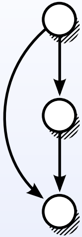
- In $O(n)$ time we compute $|N^l(v)|$ for $v \in G$ and $l \leq d$
- How useful is the truncated version?
- What about other truncated measures?

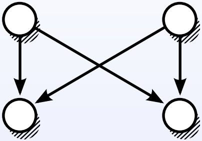


Motif/Subgraph counting

Idea: frequent structures in networks probably have a *function*

- Shen-Orr *et al.* identified network motifs in regulation network of E. coli and analyzed their function
  (Network motifs in the transcriptional regulation network of Escherichia coli. Nature Genetics 31, 2002.)
- Milo *et al.* compare network motifs of regulation networks, neural networks, food webs, electric circuits and the www
  (Network Motifs: Simple Building Blocks of Complex Networks. Science 25, 2002.)
- So far limited to motifs of size ≤ 4





Subgraph counting in bounded expansion graphs

Tool of choice: p-centered coloring.

- graph is colored with $f(p)$ colors in linear time
- every subgraph induced by $l < p$ colors has *treedepth* at most l
- Motifs of size p are colored by on of $\binom{f(p)}{p}$ color combinations

$\Rightarrow$ Problem reduced to counting in bounded-treedepth graphs!

We can do this even for disconnected graphs H in time $O(c^{|H| \log |H|} n)$ with small constants, so $\binom{f(|H|)}{|H|}$ is probably the limiting factor.

But how many colors?

Some preliminary tests: 5-centered colorings
(Can be used for patterns of size 4)

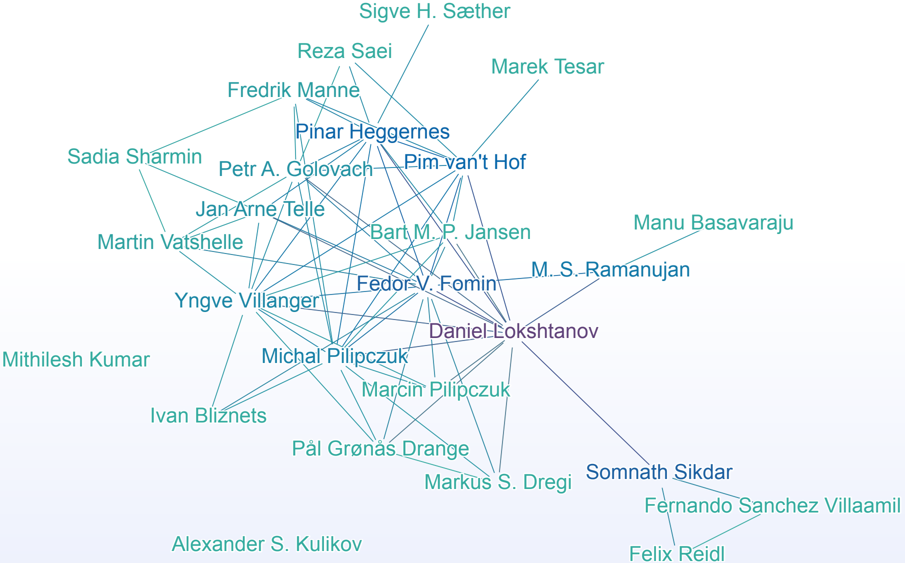
Graph	Size	Avg. deg.	Colors
netscience	1589	~ 3.5	31
diseasome	1419	~ 7.7	36
codeminer	726	~ 1.4	64
cpan-authors	840	~ 2.7	63
c. elegans	306	~ 7.7	149
football	115	~ 10	113
cpan-dist.	2719	~ 1.8	140?

Thanks to our student Kevin Jasnik for the computation!

Conclusion

- Random models of networks seem to suggest that they are graphs of bounded expansion
- A lot of algorithmic questions are open in that field
- We have some idea of how to design algorithms for this class, but it's far from settled
- Preliminary experiments show that the p-centered coloring numbers are quite low for some networks (for others not)
- We need good heuristics for these colorings!

Thanks!


Sigve H. Sæther
Reza Saei
Marek Tesar
Fredrik Manne
Pinar Heggernes
Sadia Sharmin
Petr A. Golovach
Pim van't Hof
Jan Arne Telle
Manu Basavaraju
Bart M. P. Jansen
Martin Vatshelle
M. S. Ramanujan
Fedor V. Fomin
Yngve Villanger
Daniel Lokshtanov
Mithilesh Kumar
Michal Pilipczuk
Marcin Pilipczuk
Ivan Bliznets
Pål Grønås Drange
Somnath Sikdar
Markus S. Dregi
Fernando Sanchez Villaamil
Alexander S. Kulikov
Felix Reidl

Resources

- C. Elegans image by Tormikotkas taken from http://commons.wikimedia.org/wiki/File:Caenorhabditis_elegans_Oil-Red-o.tif
- Datasets with references available at http://wiki.gephi.org/index.php/Datasets