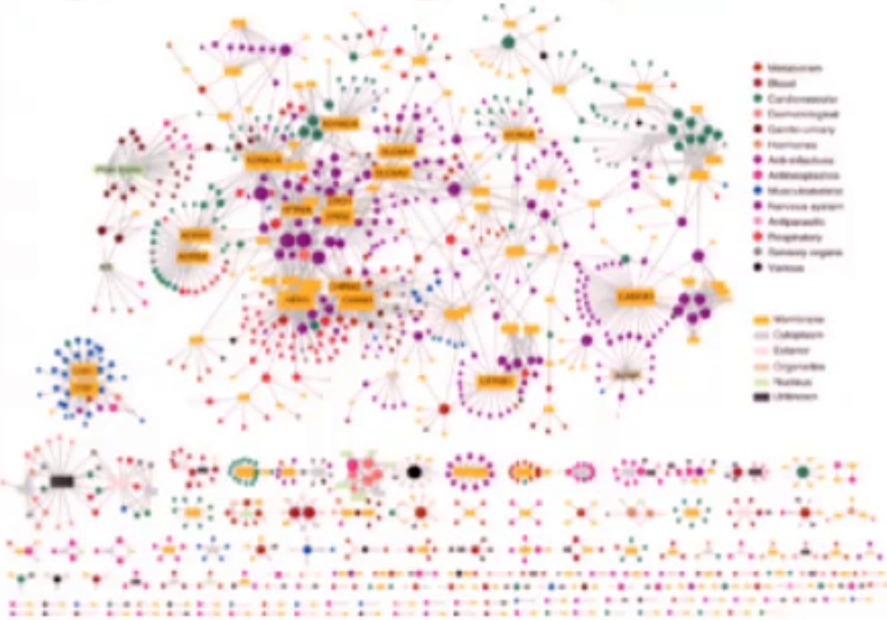
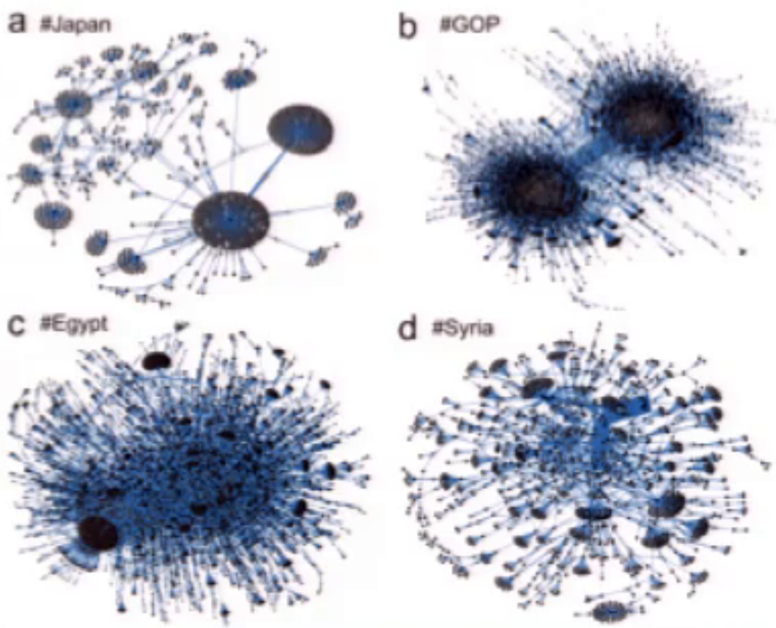
Big Data: Big Graphs





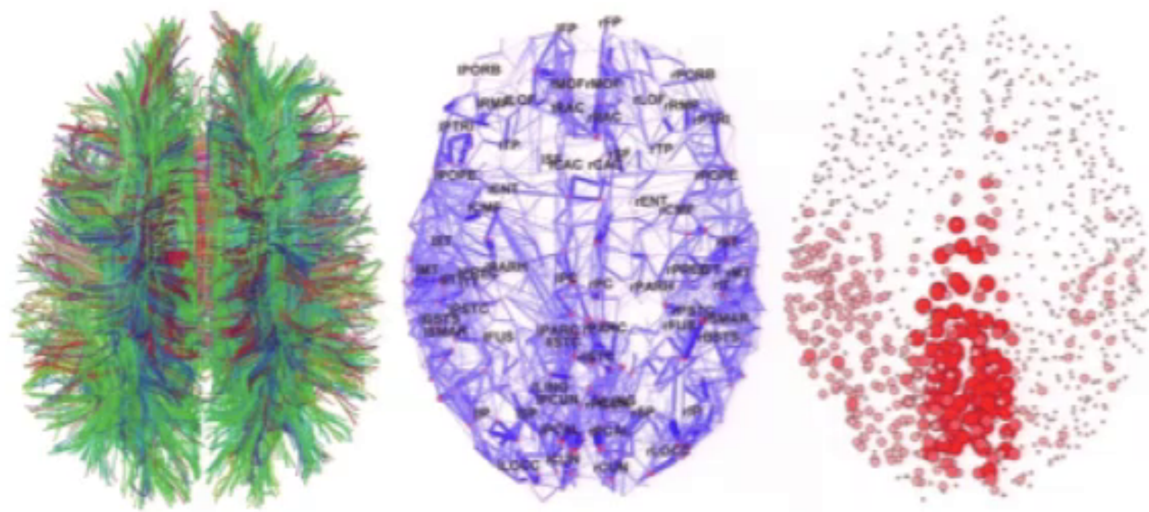


Image Credit (clockwise): Nature Biotechnology 25(10), Reidl (from Physical Review Publications: 1893–2009) , Hagmann et al., Weng et al

Fast [Graph] Algorithms?

Fact: Most interesting graph problems are NP-complete for general graphs

- Subgraph isomorphism (motifs)
- Vertex cover (sensor networks)
- Independent set (linear algebra)
- Max clique (protein groups)

"I can't find an efficient algorithm, but neither can all these famous people."

A Few "Good" Things to Limit

in a (highly debatable) increasing order of complexity

Density

Degeneracy

Hyperbolicity

Treewidth

Expansion

Bad news…

This talk is (almost) all about the definitions!

Saved by Sparsity?

Observation 1 (v. 0.0)
Many real-world networks have low average degree.

Observation 2:
Edges are not evenly distributed in real graphs.

Facebook: |V| ~ 1.3B, |E| ~ 500B
Yeast PPI: |V| = 3000, |E| ~ 3000
Power Grid: |V| ~ 5K, |E| ~ 13K
Neurome: |V| ~ 10^{10}, |E| ~ 10^{14}
Twitter: |V| ~ 1B, |E| ~ 200B

(easy to see if you think social)
Twitter: avg followers: 200, max followers ~59M
Facebook: # friends varies inter- vs intra-community

Consequences:
Some algorithms get faster (but NP-hard problems remain)

Hypothesis:
This "should" help. But how exactly?

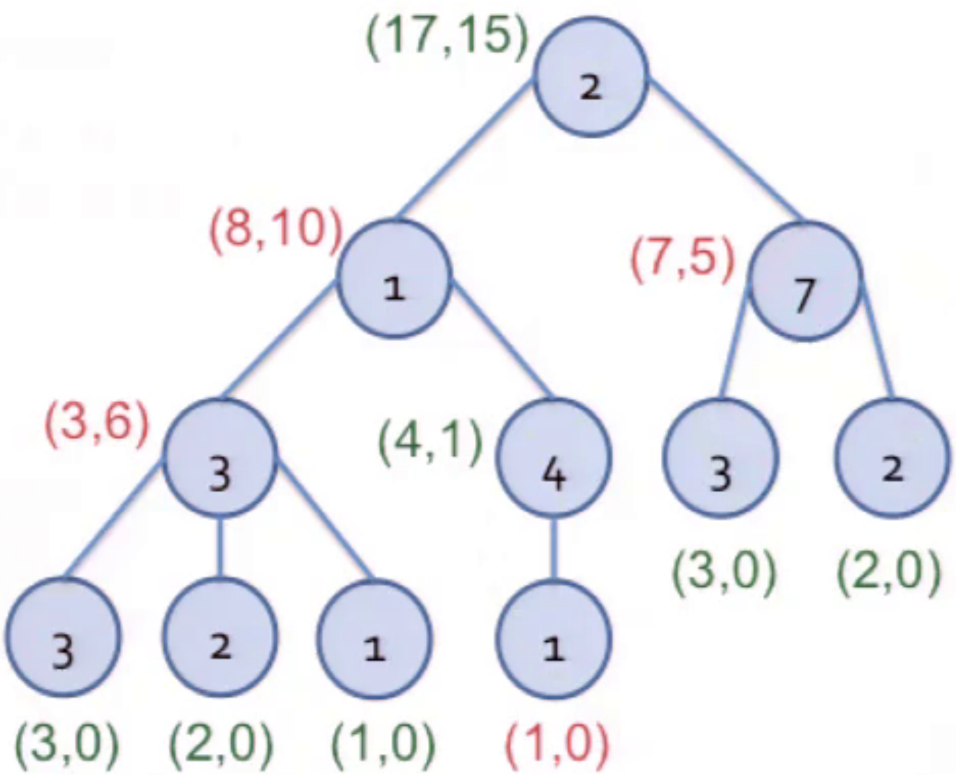
What could it mean to be structurally sparse?

Images credit Felix Reidl (RWTH Aachen)

Sparsity 1.0: Tree Structure

No cycles makes things easy!

- MaxWIS: Find the *maximum weighted independent set* in *G*



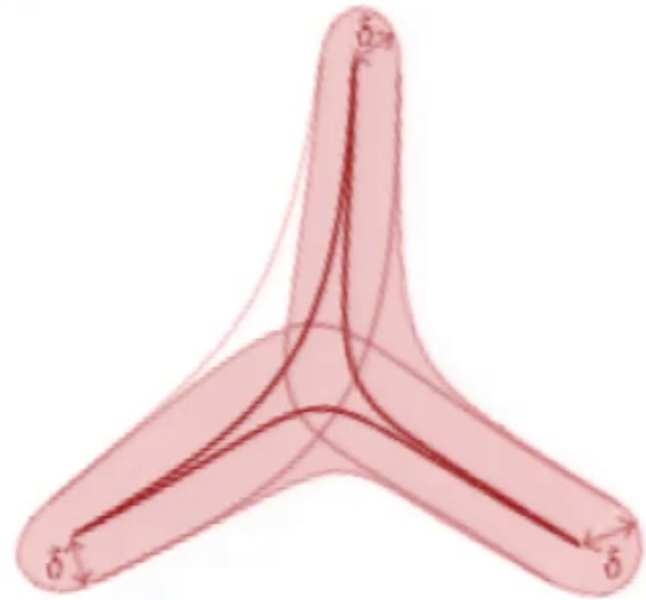
This NP-hard problem has a linear algorithm on trees!

For those who care, belief propagation also has nice algorithms on trees.

Sparsity 1.1: δ-hyperbolicity (tree-like structure 1)

δ measures the extent to which a (geodesic) metric space embeds in a tree metric [lower is better].

There are several equivalent definitions (up to constant factors): δ-slim, δ-thin, or δ-fat triangles, and Gromov's 4-point condition.
In our proofs we use the following:

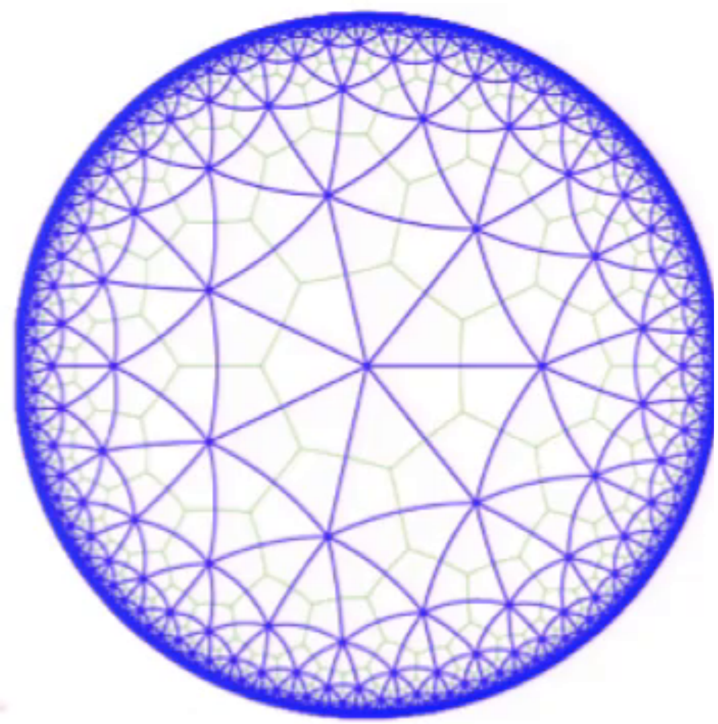


- A geodesic triangle is called **δ-slim** if each of its sides is contained in the δ -neighborhood of the union of the other two sides).
- A metric space (graph) is **δ -hyperbolic** if all its geodesic triangles are δ -thin (or δ-slim); each results in a slightly different min δ, related to each other by small constant factors.

Image credit: Bridson, Haefliger

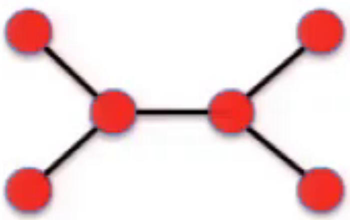
Aside: Hyperbolic Space

- Multiple parallel lines pass through a point, and angles in a triangle sum to < 180.
- Hyperbolic space gives us "extra room" to embed networks (as opposed to Euclidean space).
- In Euclidean space, a circle's area grows polynomially with its diameter; in hyperbolic space, it grows exponentially.
- Shortest paths in hyperbolic spaces are arcs through disk, not paths around the exterior.

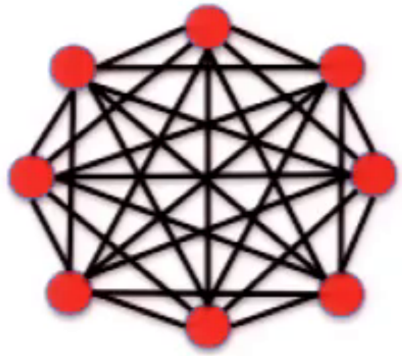


Images credit: M. Escher, Krioukov et al.

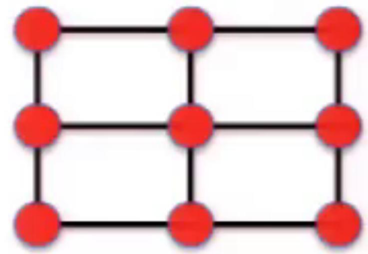
Examples and Implications



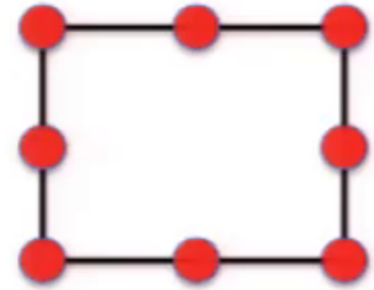
δ = 0



δ = 0



δ = √n-1



δ = n/4

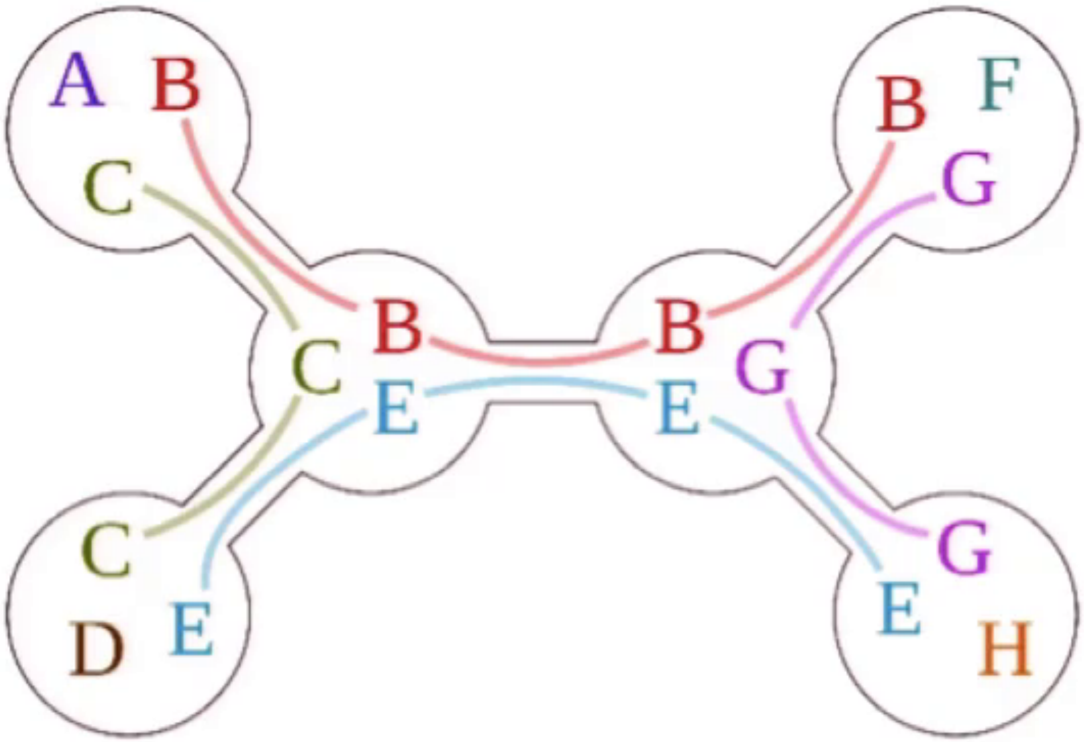
Warning: Low hyperbolicity doesn't imply traditional sparsity!

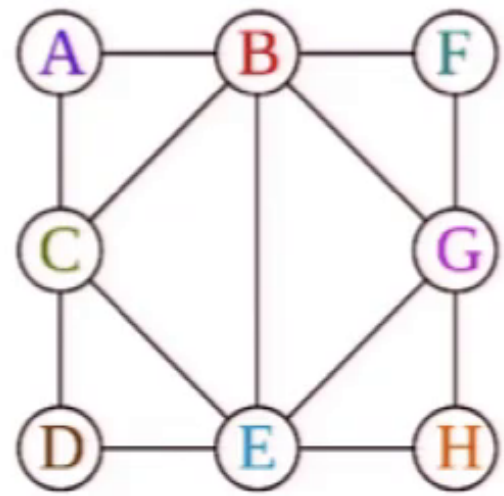
Algorithms for graph classes of bounded hyperbolicity often exploit computable approximate distance trees (Chepoi et al) or greedy routing (Kleinberg).

Work of Narayan/Saniee and Jonckheere et al conjectures that some of the observed **congestion** in real-world networks may be due to their negative curvature (hyperbolicity).

Sparsity 1.2: Bounded Treewidth (tree-like structure 2)

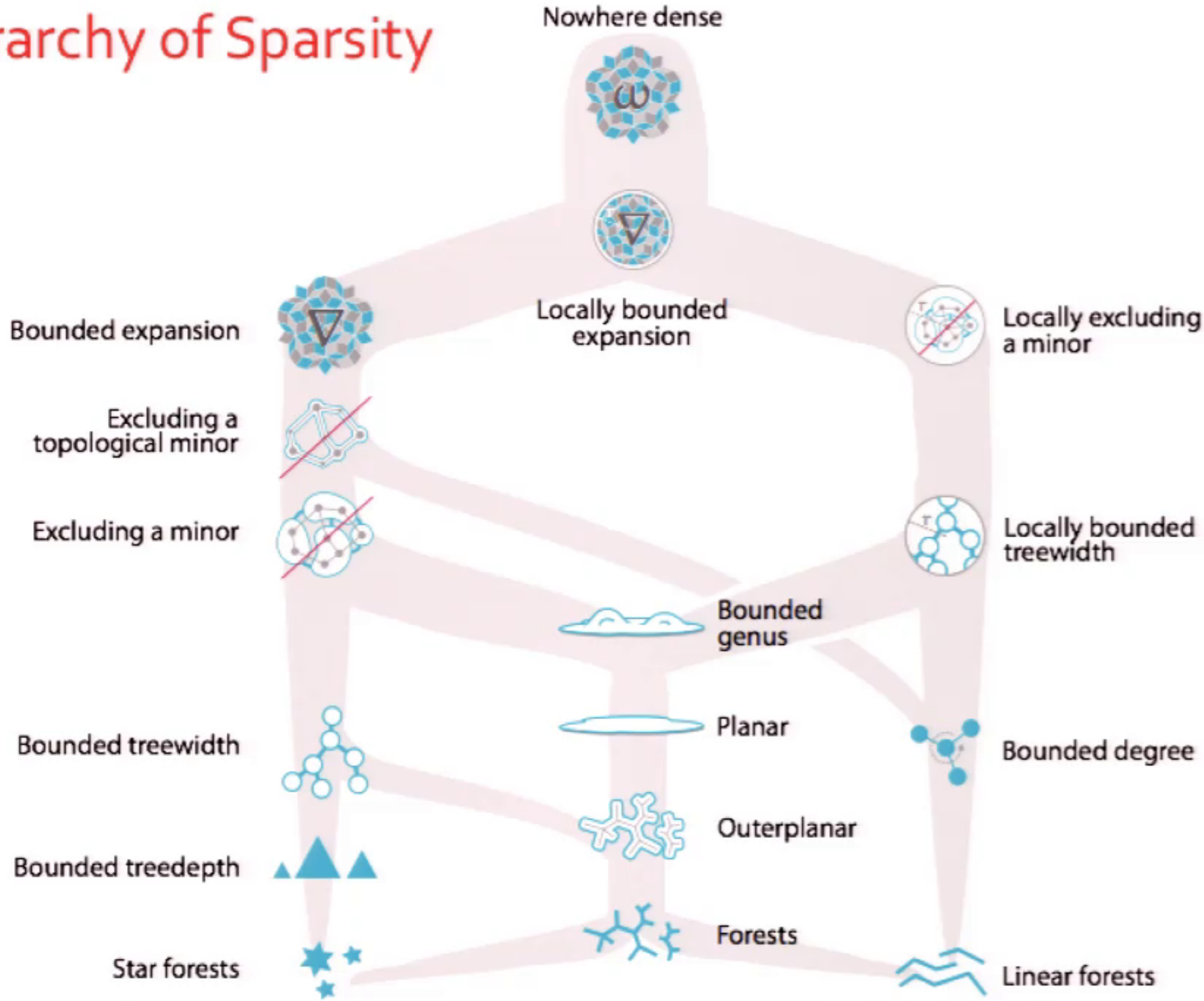
- A graph class *G* has *bounded treewidth* if every graph has a *tree decomposition* of width at most *c*.





Usefulness: Most algorithms which are polynomial-time on trees can be extended to work in poly-time on bounded treewidth (pay an exponential factor in terms of the width)

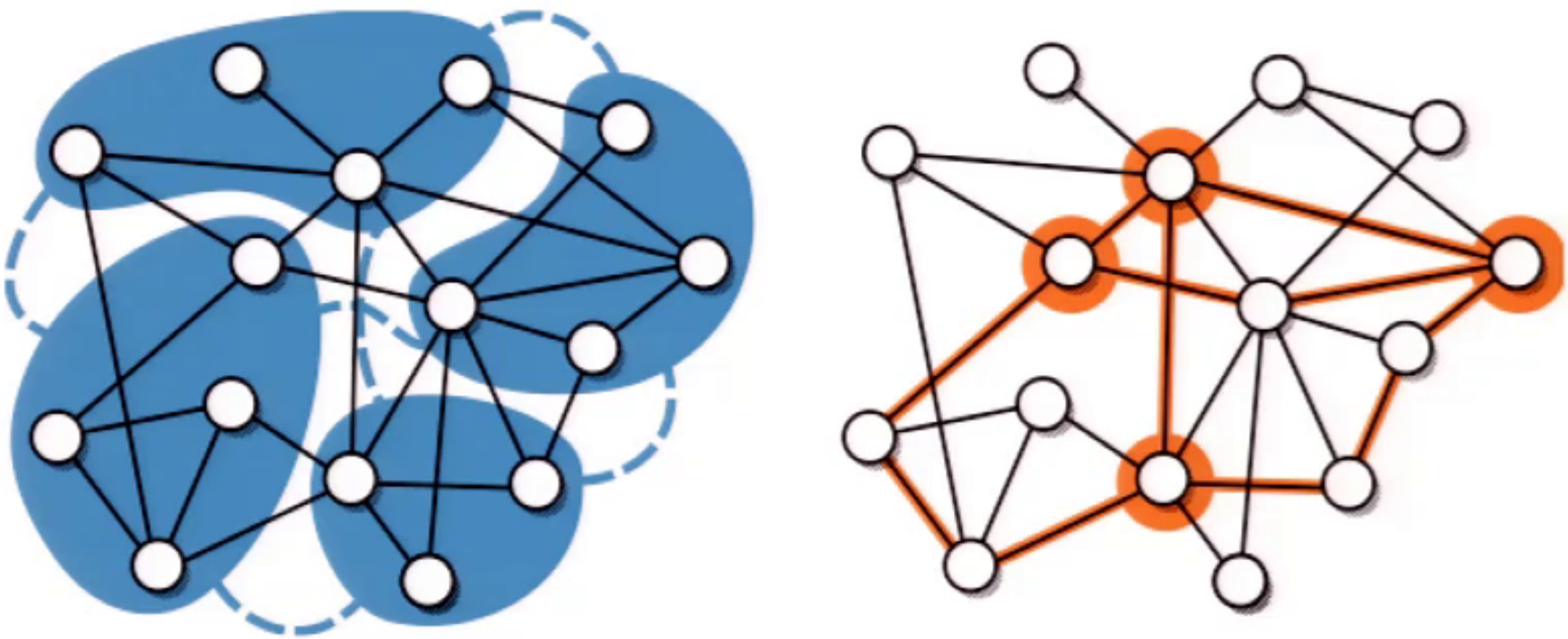
A Hierarchy of Sparsity



Sparsity 3.0: Bounded Expansion

- A graph class *G* has *bounded expansion* if every *r*-shallow (topological) minor has density at most *f(r)*.

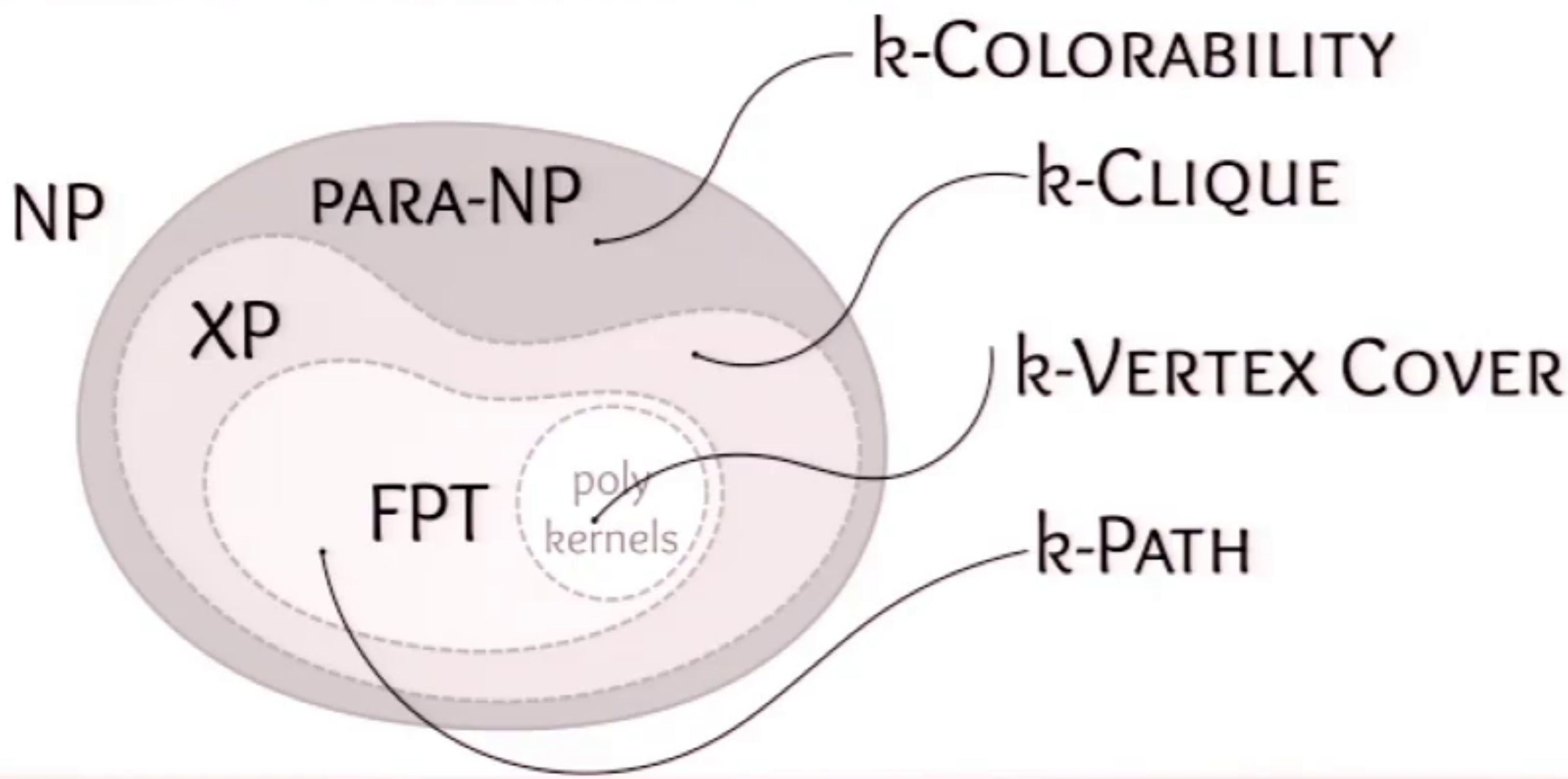
$$\nabla_r(G) = \max_{H \in G \nabla r} \frac{|E(H)|}{|V(H)|} \qquad \nabla_r(\mathcal{G}) := \sup_{G \in \mathcal{G}} \nabla_r(G) \leq f(r)$$



Note: Algorithms don't require knowing *f(r)*.

Aside: Parameterized Complexity

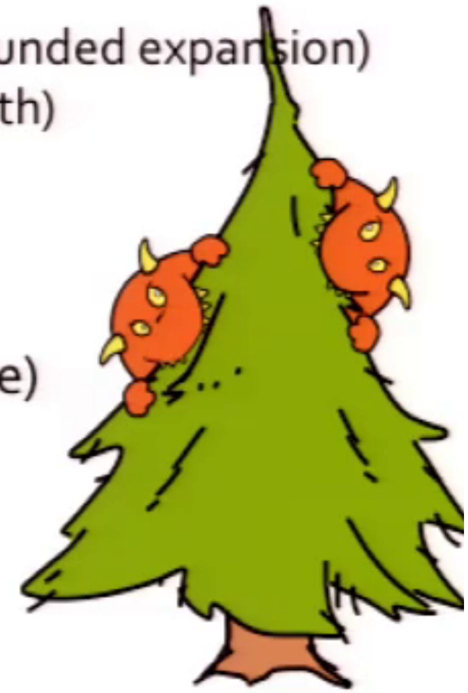
- **NP**: solvable with non-deterministic Turing machine
- **XP**: has an $O(n^{f(k)})$ algorithm.
- **FPT**: has an $f(k)n^{O(1)}$ algorithm.
- **Poly-kernel**: the kernel has size $k^{O(1)}$



FPT Algorithms & Graph Structure: Pros & Cons

Lots of problems become FPT:

- Steiner Tree (bounded degeneracy)
- DominatingSet (bounded genus)
- SubgraphIsomorphism (bounded expansion)
- MaxWIS (bounded treewidth)



And there are meta-theorems!

- FO-model checking *on nowhere-dense graphs** (k = formula size)
- EMSO-model checking *parameterized by treewidth*

BUT

Real-world networks might not fall into any of these categories!

Worse, it's hard to test membership & many existing results are negative

The algorithms often have [huge] hidden constants

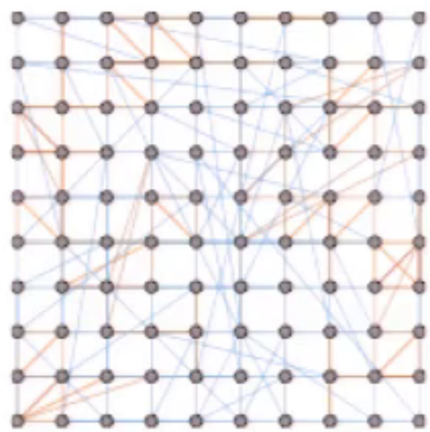
*Recall, this was broadest class. And lots of problems are expressible in FO-logic.

How do we know if real networks have these fancy variants of sparsity?

Connecting (with) the dots

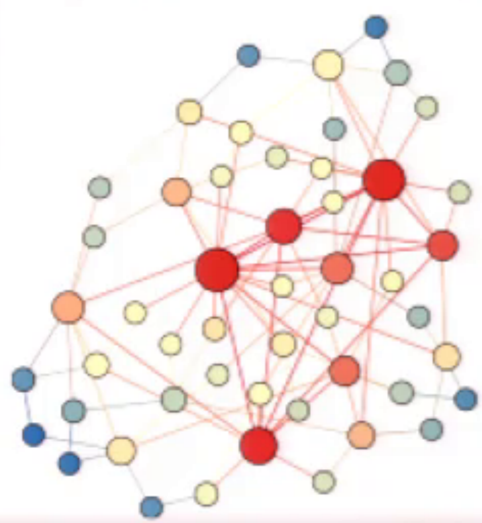
- Challenge: *instances vs. classes*
- Goal: classify networks by their features/characteristics
- Typical Approach: find *randomized models* that match desired features – use these to represent the class

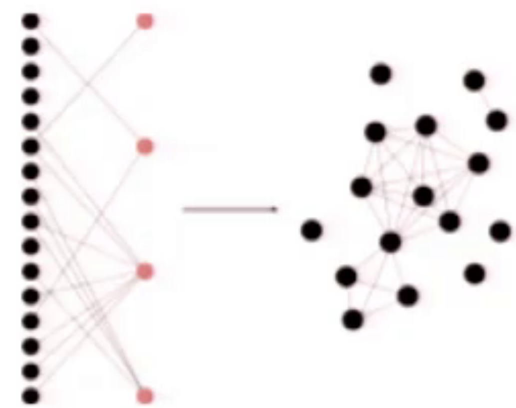
Configuration/ Chung-Lu (degree distribution)



Kleinberg (small world)

Barabasi-Albert (pref. attachment)





Random Intersection Graphs (shared attributes)

A bit of bad news

Bounded Edge Density:

- Holds in practice, but doesn't speed up many algorithms

Bounded Degree:

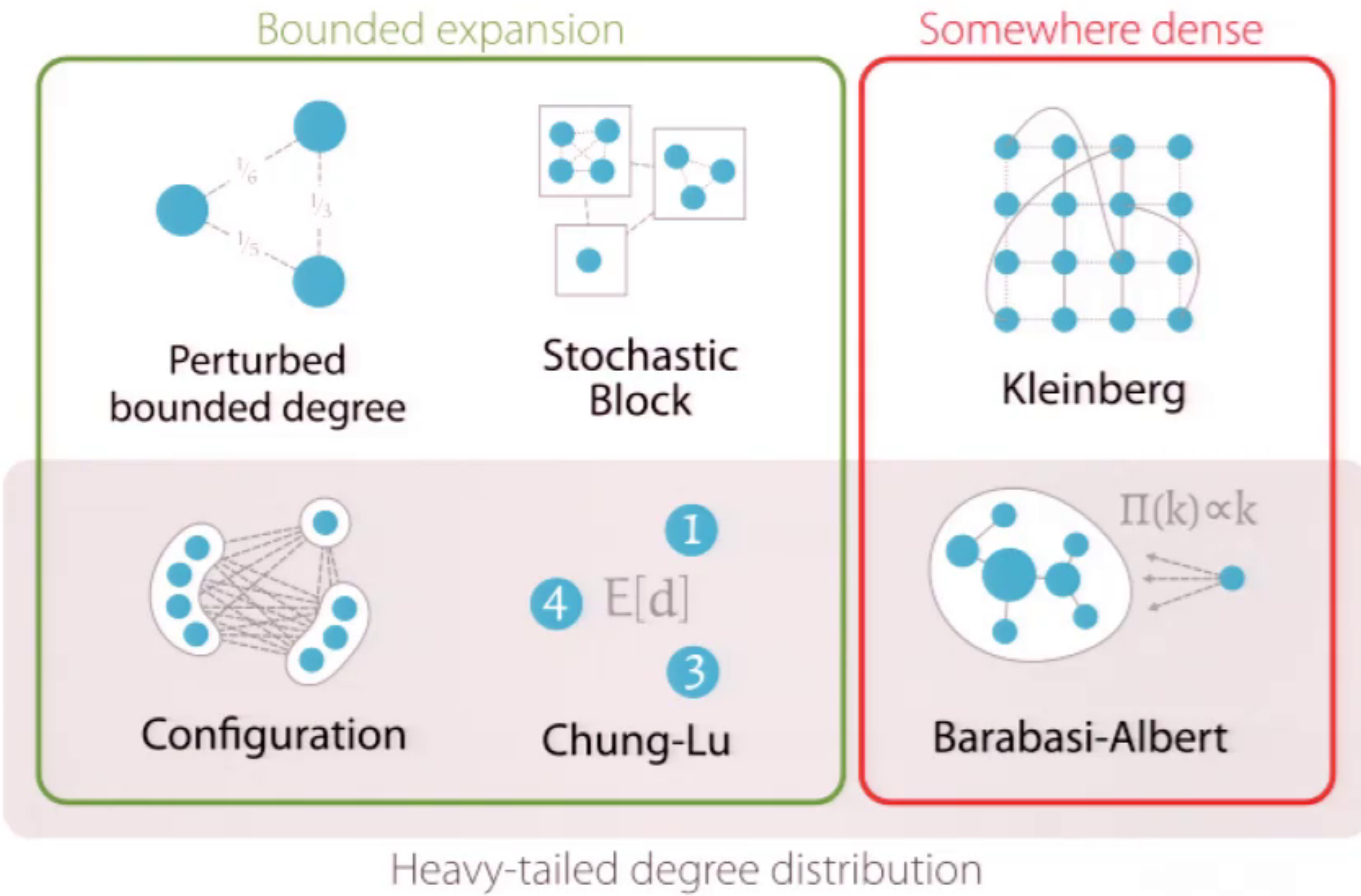
- Erdös-Rényi - O(log n)
- Molloy-Reed [depends on specified degree sequence]
- Random Intersection Graphs ($\alpha \leq 1$)

Bounded Treewidth:

- Erdös-Rényi - O(n) [Gao, 2009]
- Barabasi-Albert – O(n) [Gao, 2009]
- Empirical evidence on real data [Adcock et al, 2013]

Related work: A Plethora of New Classifications!

[with Demaine, Reidl, Rossmanith, Sanchez Villaamil, Sikdar; 2015+]



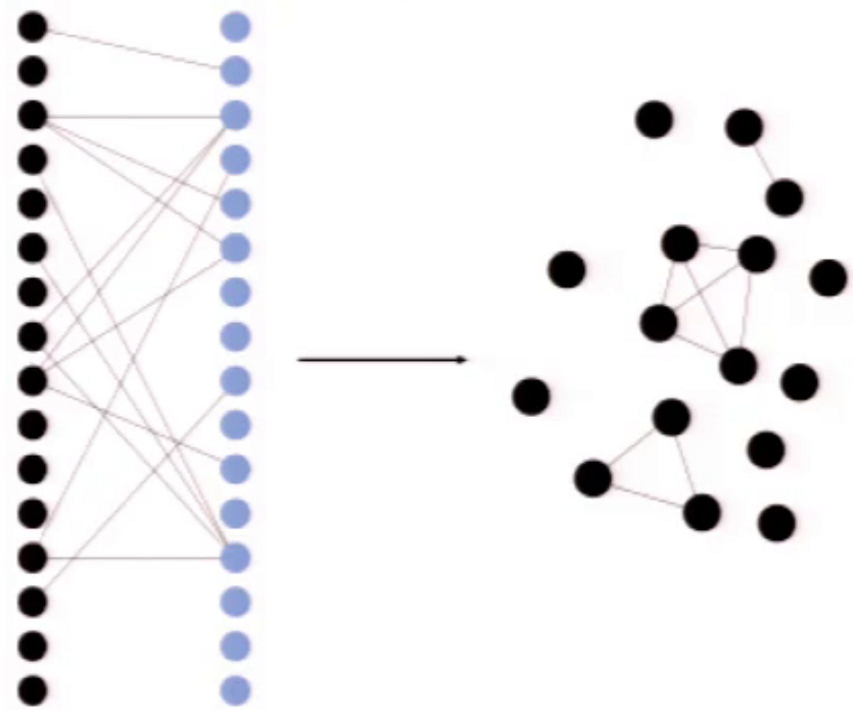
* Includes configuration with households (high clustering) & inhomogenous random graphs.

Today's focus: Random Intersection Graphs

- Introduced by Karónski, Scheinerman, Singer-Cohen in the late 90's.
- Model collaboration graphs (from arXiv), affiliation groups, etc.

- Let n be the number of nodes and α, β, γ be constants.
- Set $m = \beta n^{\alpha}$, B a bipartite graph with parts U,V of size n and m, respectively.
- For every pair u in U and v in V, add the edge (u,v) with probability $p = \gamma n^{-(1+\alpha)/2}$.
- More generally, the inhomogenous model allows p to depend on the attribute.

RIG(n,m,p) is the graph on U where u_1 and u_2 are adjacent iff there exists v in V so that (u_1,v) and (u_2,v) are edges of B.

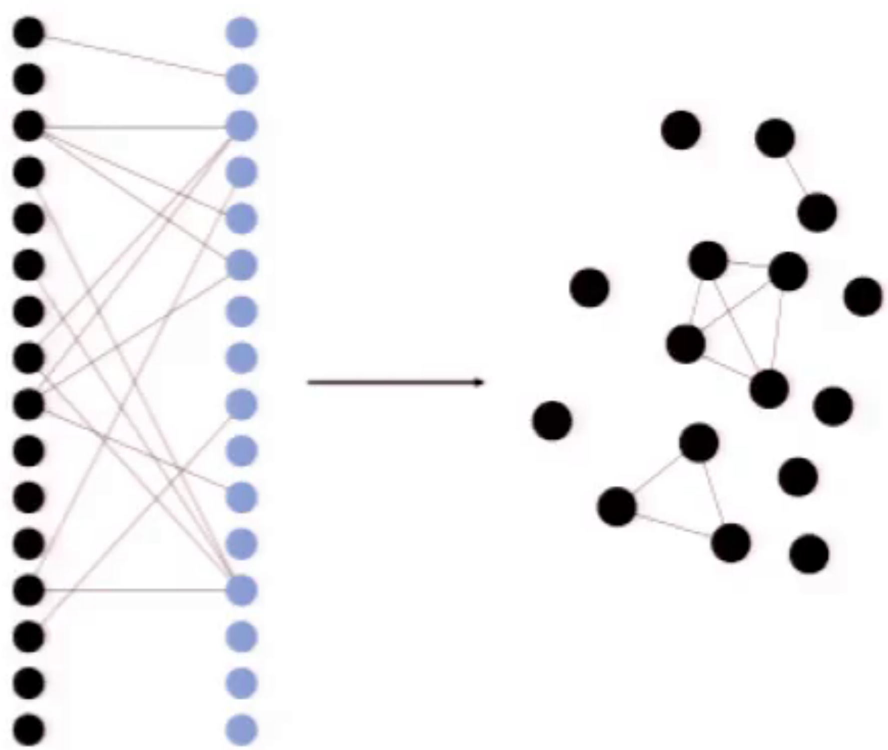


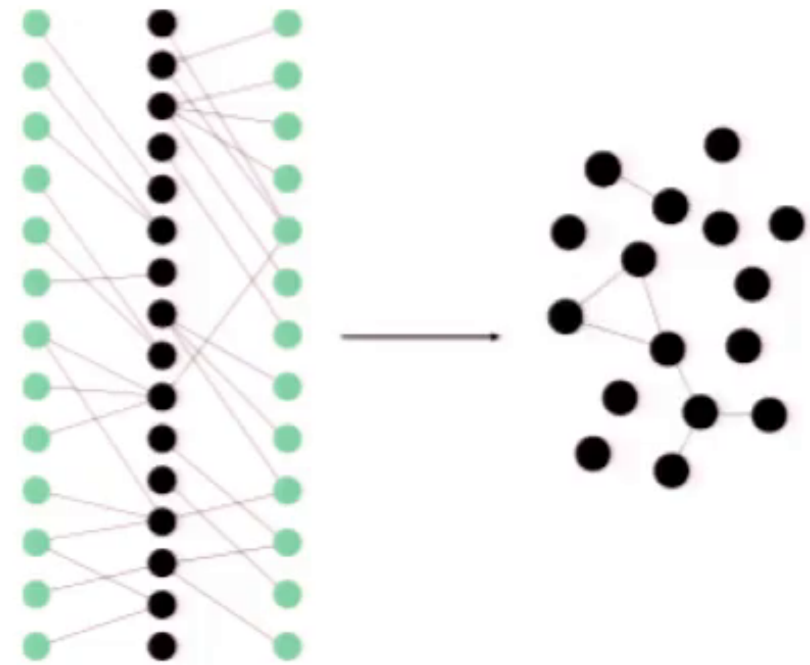
e.g. Newman's Network Science

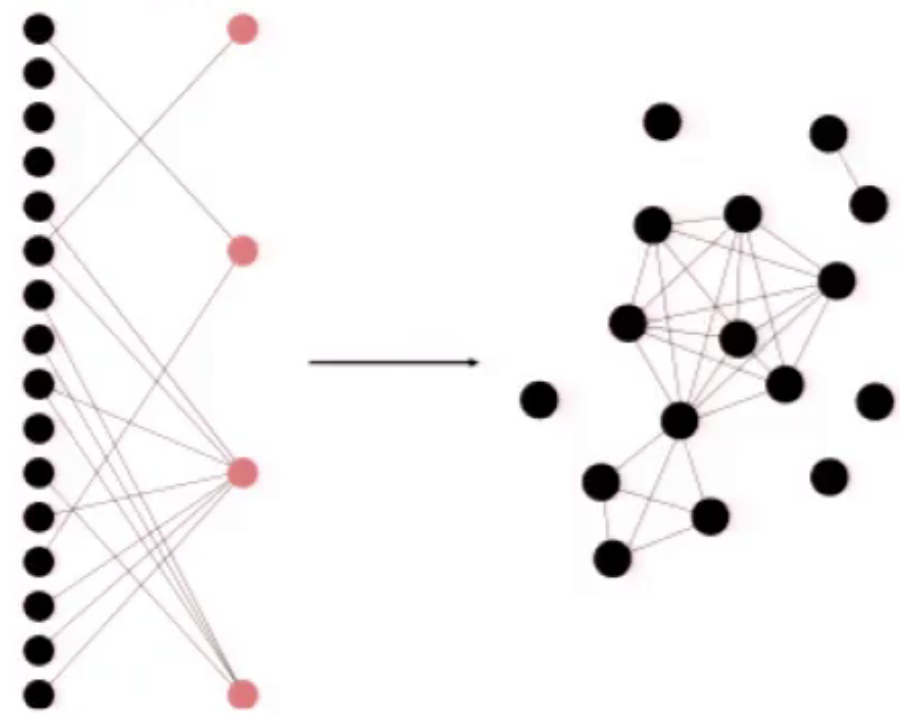
Drawn from Physical Review Publications: 1893–2009

Random Intersection Graphs

- $m = \beta n^{\alpha}$ creates three regimes of behavior: $\alpha < 1$, $\alpha = 1$, $\alpha > 1$







RIG Results!

Degeneracy & Expansion

Theorem: RIG(n, m, p) has *degeneracy*:

$\Omega(\gamma n^{(1-\alpha)/2})$ when $\alpha < 1$

$\Omega(\log n / \log \log n)$ when $\alpha = 1$ and

$O(1)$ when $\alpha > 1$.

Theorem: For $\alpha \leq 1$, w.h.p. RIG(n, m, p) is *somewhere dense* (contains arbitrarily large cliques as shallow minors), and thus not bounded-expansion.

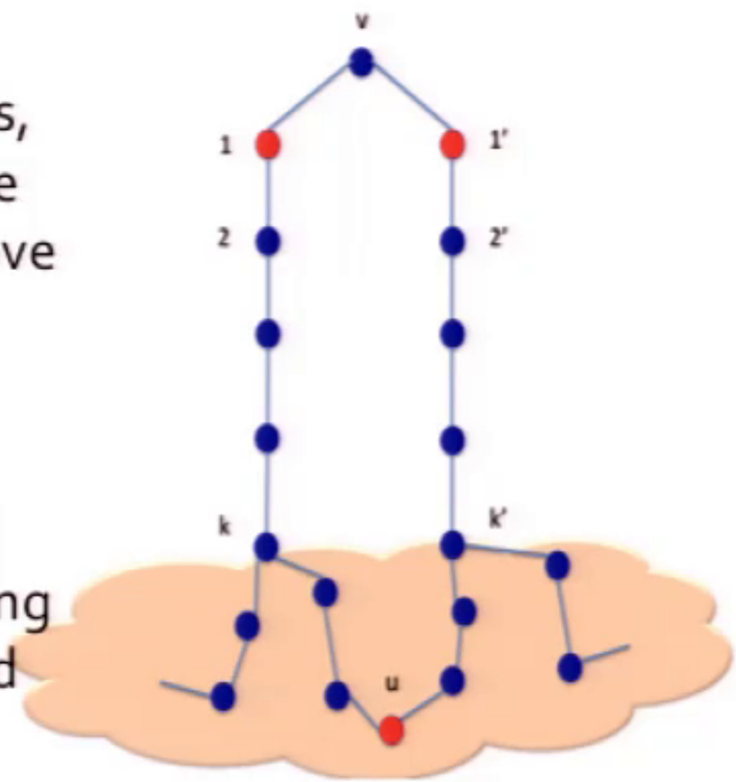
Theorem: For $\alpha > 1$, w.h.p. RIG(n,m,p) has *bounded expansion*.

All results on this page were proved w.h.p. (**with high probability**):
for any $c \geq 1$ the event occurs with probability at least $1 - f(c)/n^c$ for large enough *n*.

(non-) Hyperbolicity

Theorem: Under reasonable restrictions on β and γ, a.a.s. RIG(n, m, p) has hyperbolicity Ω(log n) for all values of α.

Proof Sketch:
We extend the method of Narayan et al for ER graphs, and randomly "expose" a large enough fraction of the vertices to w.h.p. contain a giant component. We prove there is an induced path in the "hidden" portion of length proportional to log *n* whose internal vertices have no other neighbors in the graph and whose endpoints lie in the giant component of the exposed graph. It then follows that there is a cycle formed using this "handle" (path) which cannot have shortcuts, and thus delta is at least *k*/4.



Note this is "only" a.a.s. (**asymptotically almost surely**): probability of event tends to one in the limit.

Shameless Plug