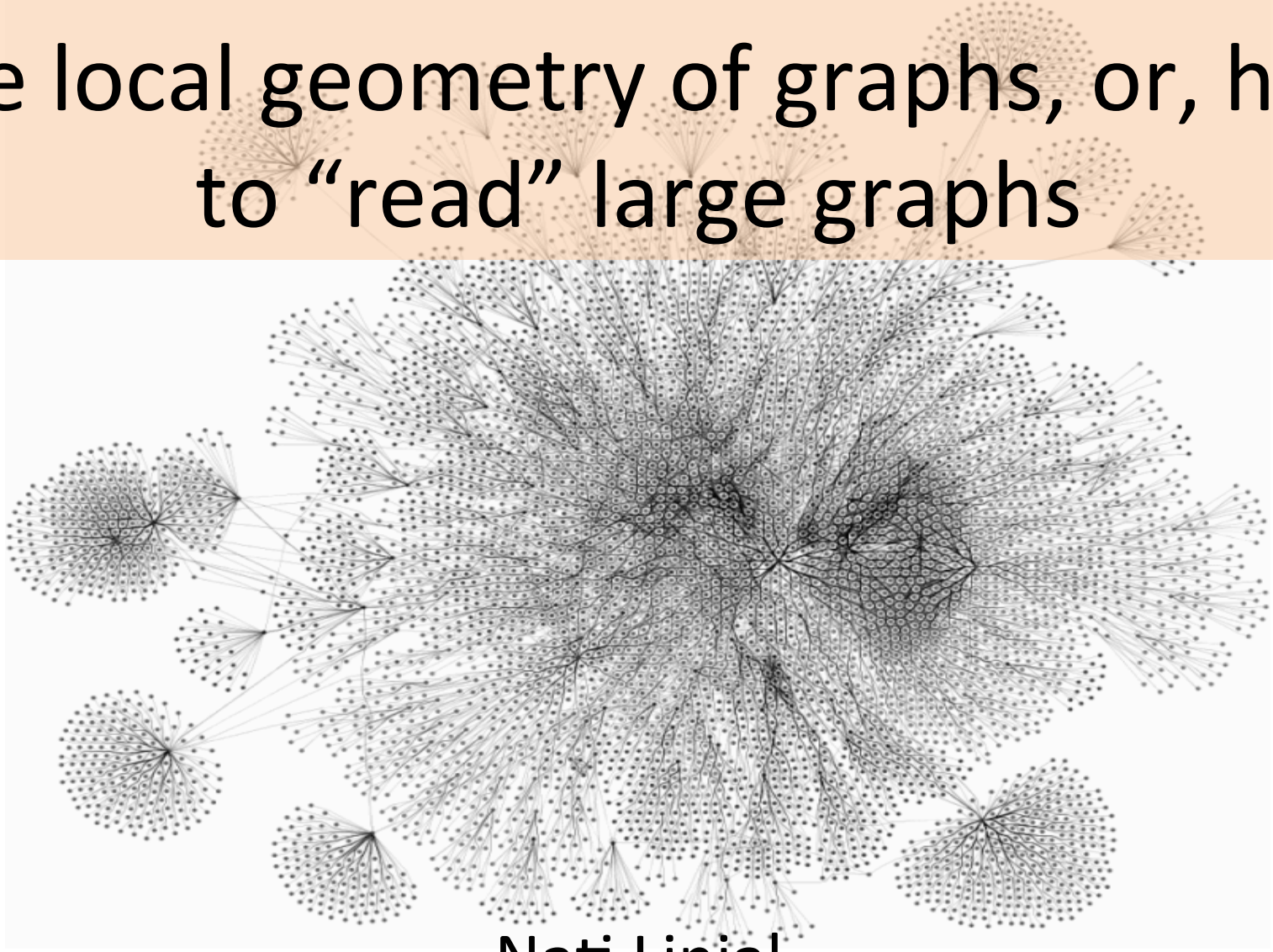
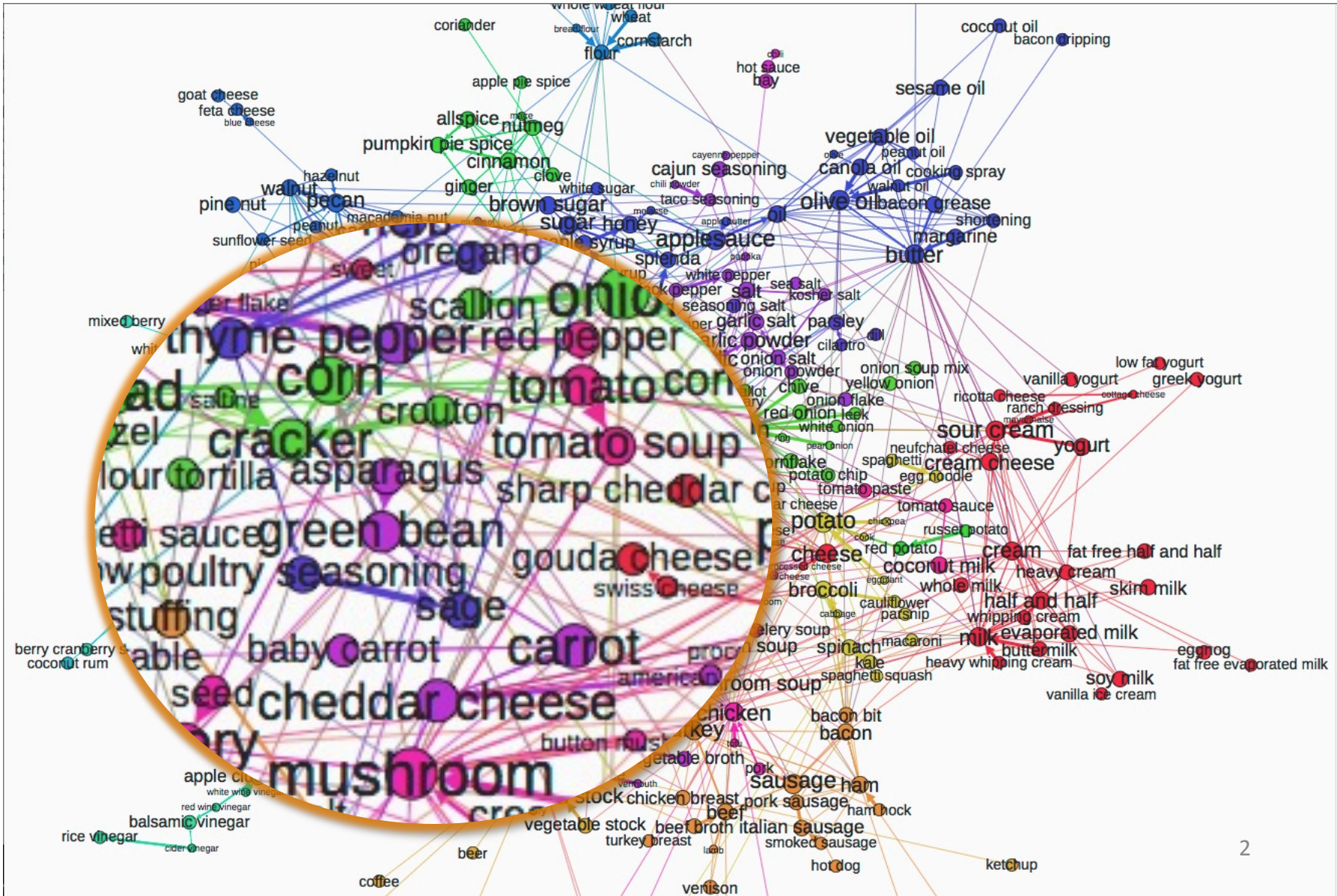
The local geometry of graphs, or, how to “read” large graphs



Nati Linial

Hebrew University, Jerusalem

Graphs (aka Networks)



Statistics 001

- What do you do with a large collection of numbers that come from some phenomenon or a system which you study?
- The most basic answer: Draw a histogram, look at key parameters – Mean, median, standard deviation…
- Try to fit to known distributions: Normal, Poisson, etc. Estimate key parameters. Draw conclusions on the system at hand.

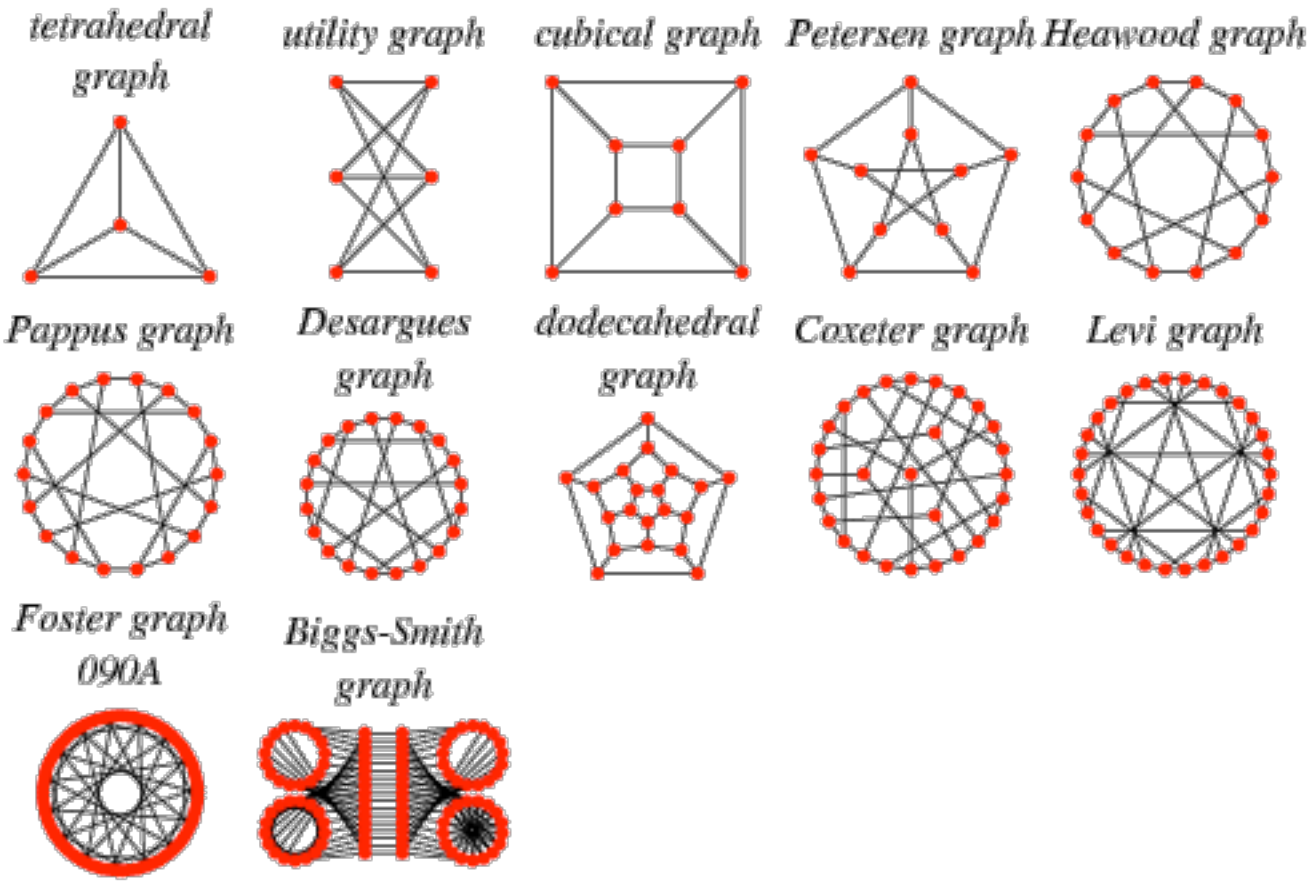
Graph reading 0.001

- We need a parallel methodology when the input is a large graph and not a big pile of numbers.
- Two necessary ingredients for this program: Find out which key parameters should be observed in a big graph (in this talk we discuss one answer to this question).
- Develop a battery of generative models of graphs and methods to recover the appropriate model from the input graph.

The main focus of this lecture

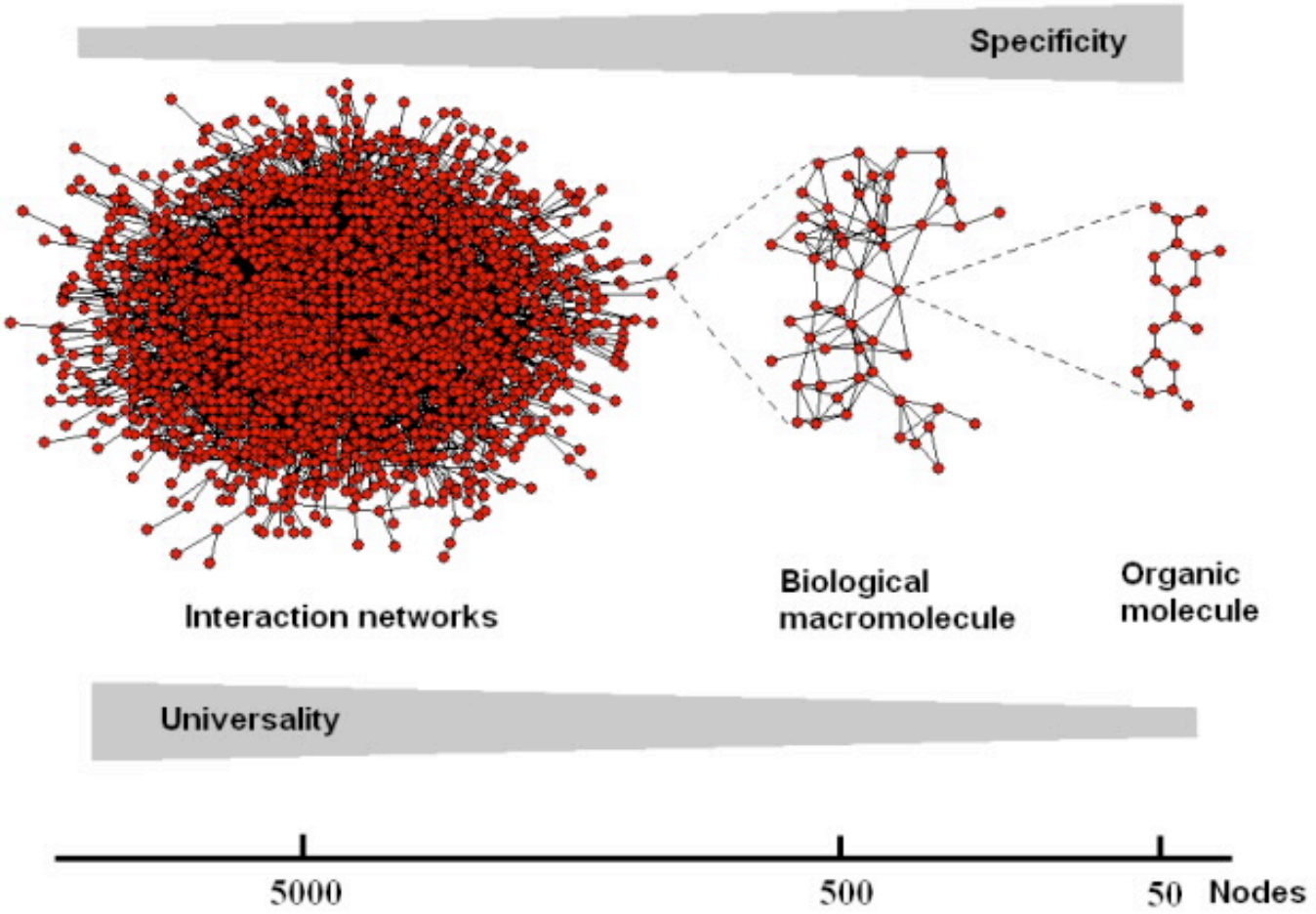
- How should we "read/understand" **very large** graphs? (Possibly graph is so big that it cannot even be stored in our computer's memory.)
- The approach that we discuss here: Sample small chunks of G (say k vertices at a time) and consider the resulting distribution on k-vertex graphs, to which we refer as a local view of G or the k-profile of G.

How do you do? Some small graphs

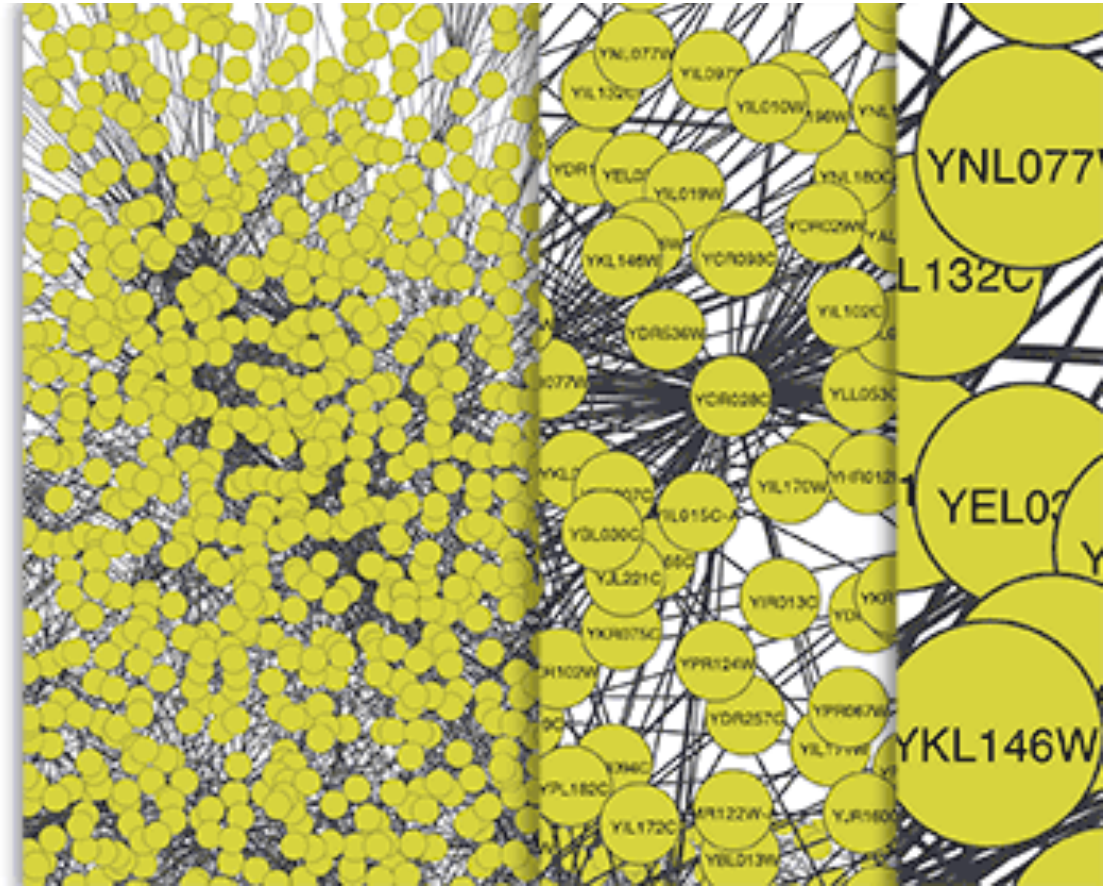


Emerging Field: Network Biology

Network Biology: The hair ball



Network Biology: The hair ball



The two major questions

- Which local views are possible? (Local graph theory). Namely, which distribution on k-vertex graphs can be obtained as the k-profile of a large graph?

- How are the global properties of G reflected in the local view? (Local-to-global theory). Namely, what large-scale structural conclusions can you infer about G, based on its local view?

Are these questions completely new?

Here are several pertinent bodies of knowledge:

- **The field of property testing**
- **Extremal and probabilistic graph theory**
- **The theory of graph limits**
- Flag algebras
- Lots of other material that we do not even touch, e.g. minor-closed families of graphs

Are these questions completely new?

Here are several pertinent bodies of knowledge:

- The field of property testing

Property testing

- We wish to determine whether a huge graph G=(V,E) has some specified graph property P.

  For example:

- Is G planar? I.e., can it be drawn in the plane so that no two edges are intersecting?
- Is it 7-colorable? I.e., can the vertices of G be colored by 7 colors so that every two adjacent vertices are colored differently?

We seek a super-fast decision method

- We insist that the computation time is bounded by a constant- Independent of the size of G, (which is assumed to be huge).
- Obviously, there are some prices to pay:

A. Our algorithm must be probabilistic, and we must allow for a chance of error.

B. Moreover, we must allow the algorithm to err on ``borderline" instances of the problem.

What is a “borderline” instance?

- Recall: We want to decide whether graph G has property P. If the answer is “yes” this is meant verbatim.
- If the answer is “no”, we only care about instances G that are “far from having property P”. I.e., in order to turn G into a graph with property P at least 1% of the pairs must be switched (neighbors ←-→ non-neighbors).

The notion of an error

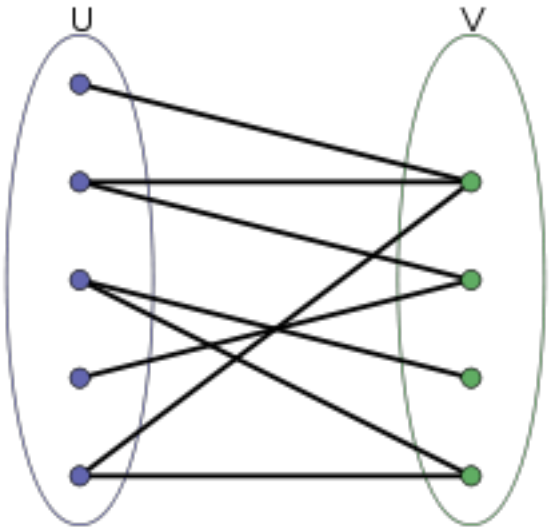
Here is what the algorithm looks like:

- Randomly sample a set of vertices S of constant size. Consider the subgraph of G induced on S. Your response depends only on this graph.
- We require a good (but possibly imperfect) success rate. Namely whatever our answer is, we must be correct with probability > ¾.

A concrete example – Is G bipartite?

- We call G bipartite if its vertices are split in two parts, say L and R and all edges connect an L-vertex to an R-vertex.
- Given access to a huge G we wish to determine whether or not it's bipartite.
- Note that a subgraph of a bipartite graph is also bipartite, hence the following algorithm suggests itself very naturally.

G bipartite?



An algorithm for testing whether a huge graph G is bipartite or not

- Randomly sample a set S of 1000 vertices in G
- Check: Is the subgraph of G induced on S bipartite or not? (This can be done efficiently)
- If it is not bipartite, respond with “G is not bipartite”. In this case you are surely correct.
- If this subgraph is bipartite declare “G is bipartite”. You are right with probability > ¾.

The crux of the matter

The last statement is quite a nontrivial theorem. It says something like:

- If a graph is 0.01-far from being bipartite, then with probability > 3/4 a randomly chosen set of 1000 vertices will reveal it.

- The mavens among you know that there is some statement with **ε** and **δ** hiding here, but we will skip such complications

In other words

- Let B and F be two huge graphs. B is Bipartite and F is 0.01-Far from being bipartite.
- Consider two distributions on 1000-vertex graphs: The one that comes from local samples of B vs. the same from F.
- The theorem says that these distributions are very different. In the B-distribution all 1000-vertx graph are bipartite, whereas in the F-distribution at most ¼ are bipartite.

In the language of the present talk

- The (global property) of being bipartite is reflected locally.
- The easy part: Every subgraph of a bipartite graph is bipartite as well.
- The hard part: In a graph that's 0.01-far from being bipartite, less than a ¼ of the 1000-vertex subgraphs are bipartite.

Something for the experts

- So, is the game over? A beautiful theorem of N. Alon and A. Shapira determines exactly which graph properties can be tested this way. Namely – hereditary graph properties.

- This answer is still far from satisfactory from the practical point of view, since the proof relies on the Szemeredi Regularity Lemma which gives a terrible dependency of δ on ε

Are these questions completely new?

Here are several pertinent bodies of knowledge:

- The field of property testing
- Extremal and probabilistic graph theory
- The theory of graph limits
- Flag algebras
- Other material that we do not go into, e.g. minor-closed families of graphs

Extremal graph theory - A parent of local graph theory

- A very intuitive thought: A graph with many edges must contain dense sets of vertices.

Extremal graph theory - A parent of local graph theory

- A very intuitive thought: A graph with many edges must contain dense sets of vertices.
- The first example: Mantel's Theorem 1907. A graph with more than $\frac{n^2}{4}$ edges (i.e., density > ½) must contain a triangle. The bound is tight.

The grandfather of extremal graph theory

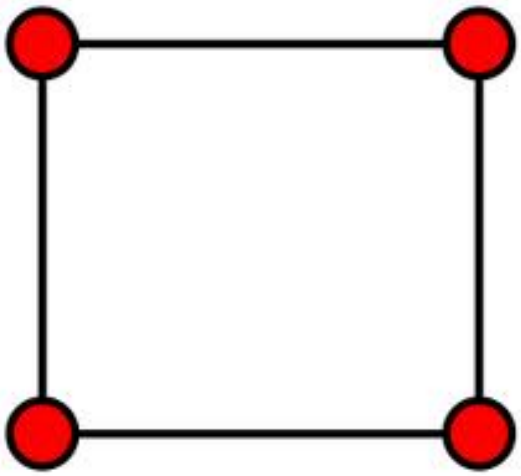
- Turan's Theorem 1941: A graph with density >(r-2)/(r-1) must contain a complete graph on r vertices. The bound is tight.

The density of large H-free graphs

- Q: Given a graph H, what is the maximum density of a large graph that does not contain a (not necessarily induced) copy of H as a subgraph.
- A: (r-2)/(r-1), where r is H's chromatic number.
- In particular, we know the answer quite accurately, unless H is bipartite (this is the case r=2 in the above).

One success with a bipartite H – The case of the 4-cycle

- The largest number of edges in an n-vertex graph that contains no 4-cycle (whether induced or not) is $\frac{n^{3/2}}{2}$



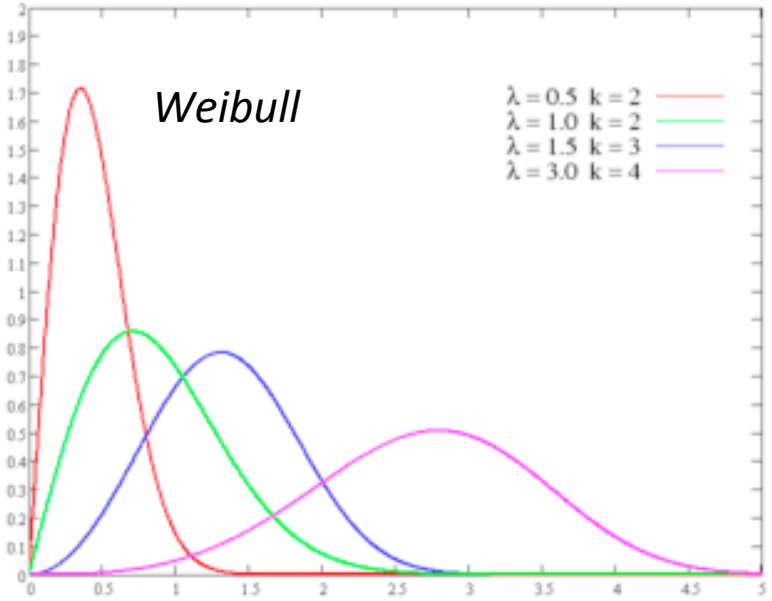
Back to “How to read large graphs ?”

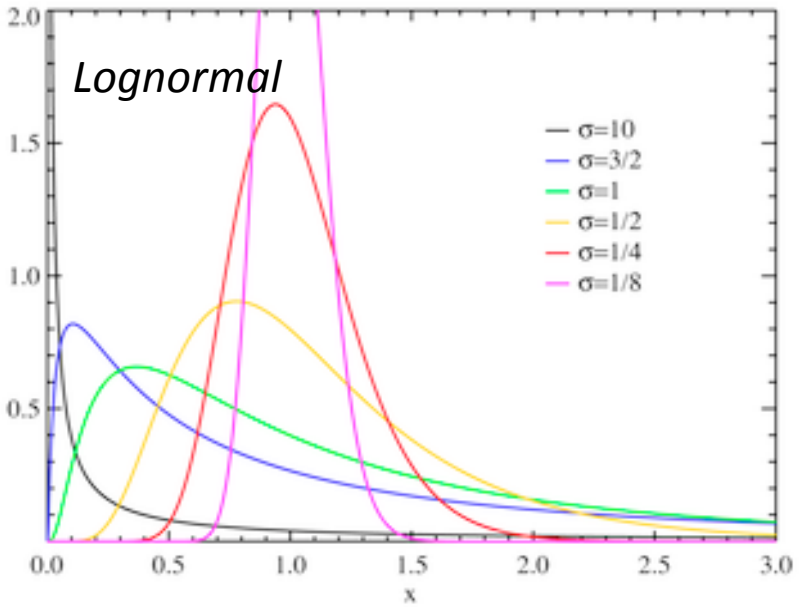
- In statistics we see a bunch of real numbers and we wish to say something worthwhile on the domain from which these numbers came.

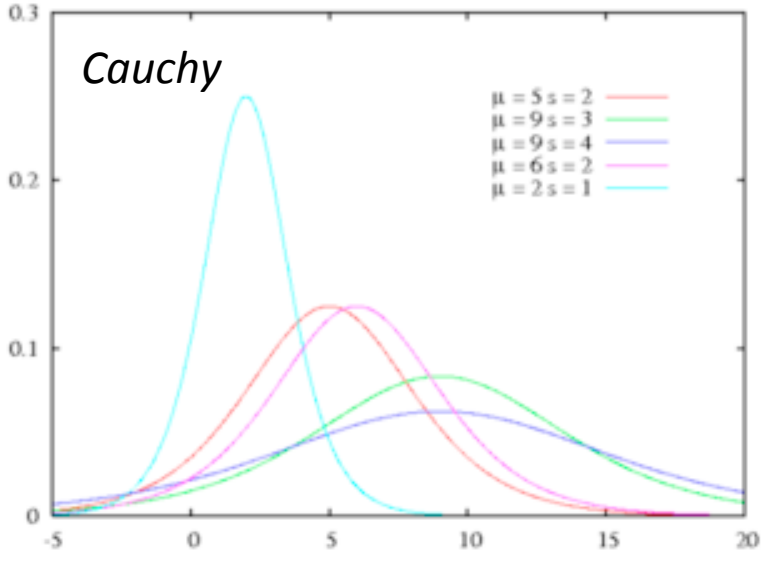
Back to “How to read large graphs ?”

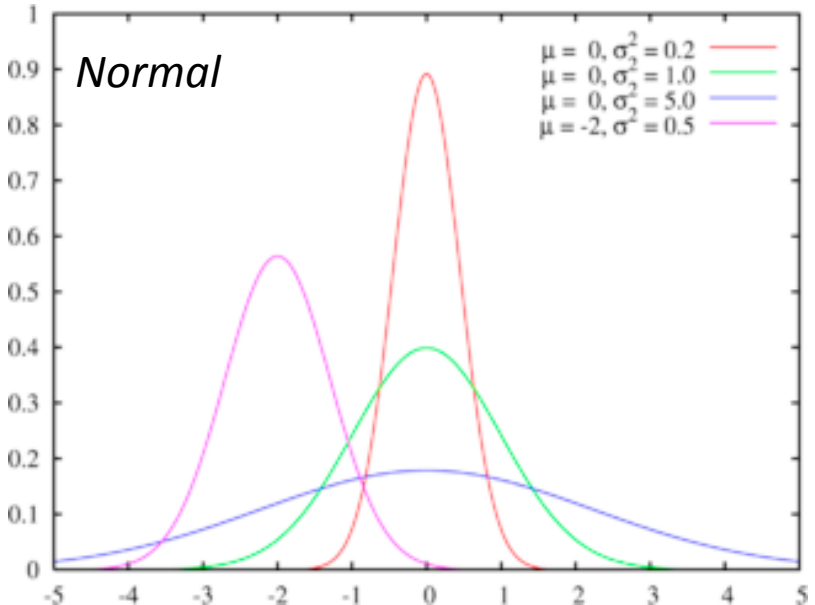
- In statistics we see a bunch of real numbers and we wish to say something worthwhile on the domain from which these numbers came.
- We realize that the (“empirical”) distribution of the given sample resembles a known distribution (e.g., Normal, Poisson, Gamma…). We estimate the relevant parameters and try to associate with the relevant domain.

Library of distributions









An analog paradigm for graphs

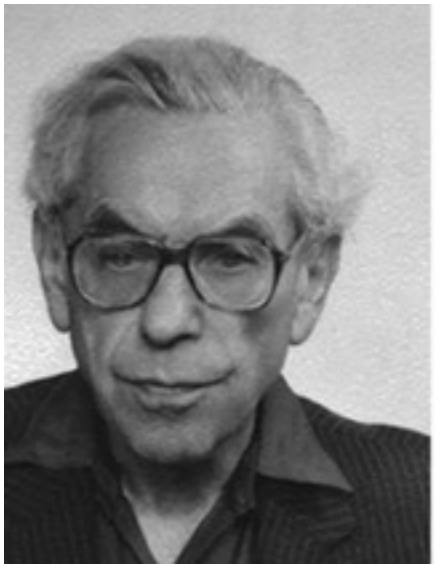
In order to "read" a large graph G, we:

1. Consider models for generating graphs.
2. Find the best fit among these models.
3. Estimate the relevant parameters.
4. Draw conclusions on the source of the data.

We seek to develop the infrastructure that's needed to make this methodology work.

Probabilistic and generative graph models

- The oldest such models go back more than 50 years, namely the Erdos-Renyi G(n,p) model of random graphs.




The G(n,p) model

- Here n is an integer (which we normally take to be large – We are interested in the asymptotic theory) and the parameter 1>p>0.
- We start with n vertices. Independently, for each pair of vertices xy we put in the edge xy.

G(n,p) theory

- This is the simplest, most basic and most thoroughly understood theory of random graphs. Very flexible and easy to investigate.
- It taught us many previously unexpected things about large graphs.
- On the other hand it's very simplistic, and too restricted for the purpose of modeling large real-life networks.

Other models of random graphs

- Random d-regular graphs. Every vertex has exactly d neighbors. “The configuration model” – Substantially different from G(n,p).

B. Bollobas

Other models of random graphs

- Percolation models – Start from a d-dimensional grid, maintain edges independently with probability p. Originated in statistical mechanics
- Random graph covers (aka “random lifts”). A model of graph that combines deterministic with stochastic ingredients.

Generative models

- Preferential attachment models: An evolving graph model. Start with a seed graph. At each step add a new vertex that becomes a neighbor of a random subset of the earlier vertices with a preference towards high-degree vertices.
- Models of growth + mutations. E.g., a random vertex spawns a ``clone" that slightly ``mutates" the neighbor set of the original vertex.

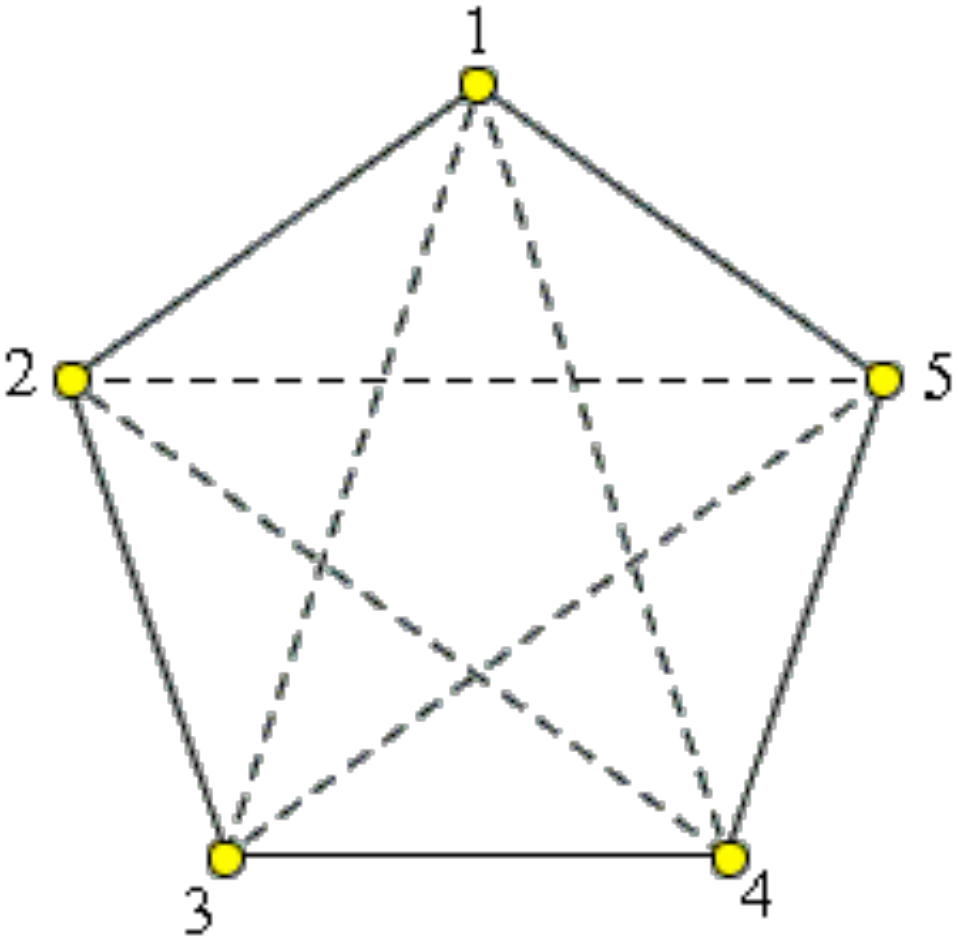
Back to local graph theory: Ramsey's theorem

- "Total chaos is impossible". This is a fundamental principle in combinatorics and in many other mathematical areas.
- In particular, every large graph must contain a substantially large homogeneous set, i.e., a clique (a subgraph in which every two vertices are adjacent) or an anti-clique (a set of vertices with no edges).

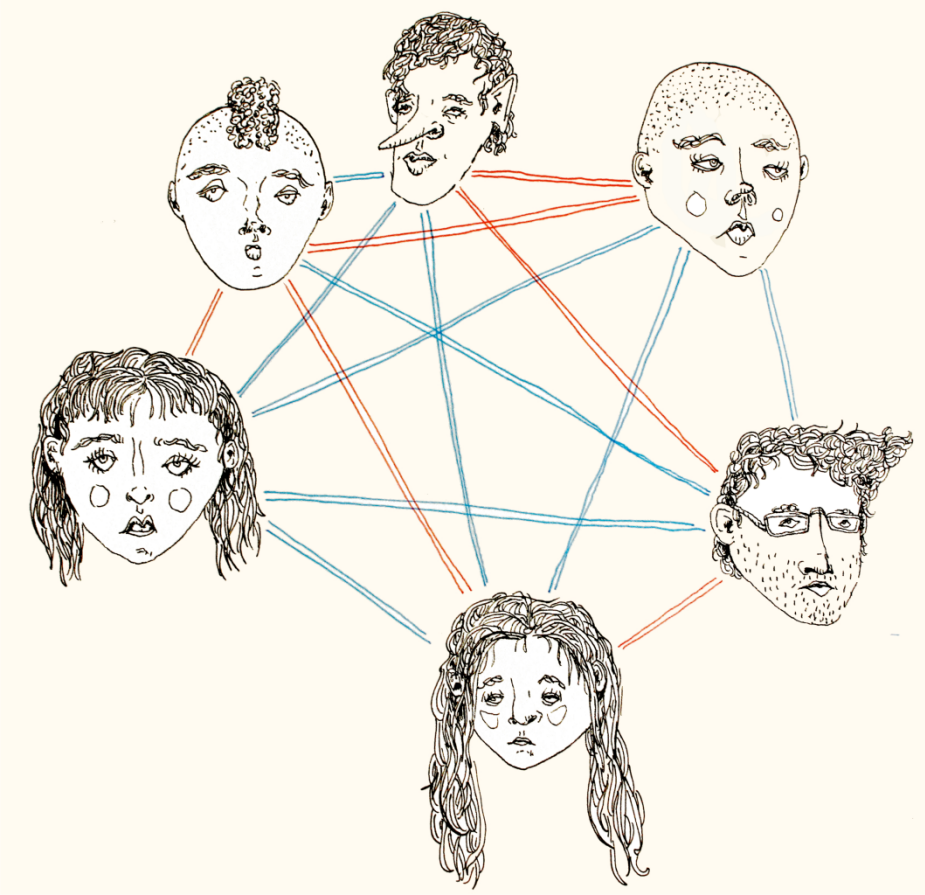
Quantitative Ramsey Theorems

- In a party of 6 people there are 3 people who are mutually acquainted or 3 who are mutually unacquainted.
- But this need not be the case in a party of 5.
- In a party of 18 people there are 4 people who are mutually acquainted or 4 who are mutually unacquainted.
- But this need not be the case in a party of 17.

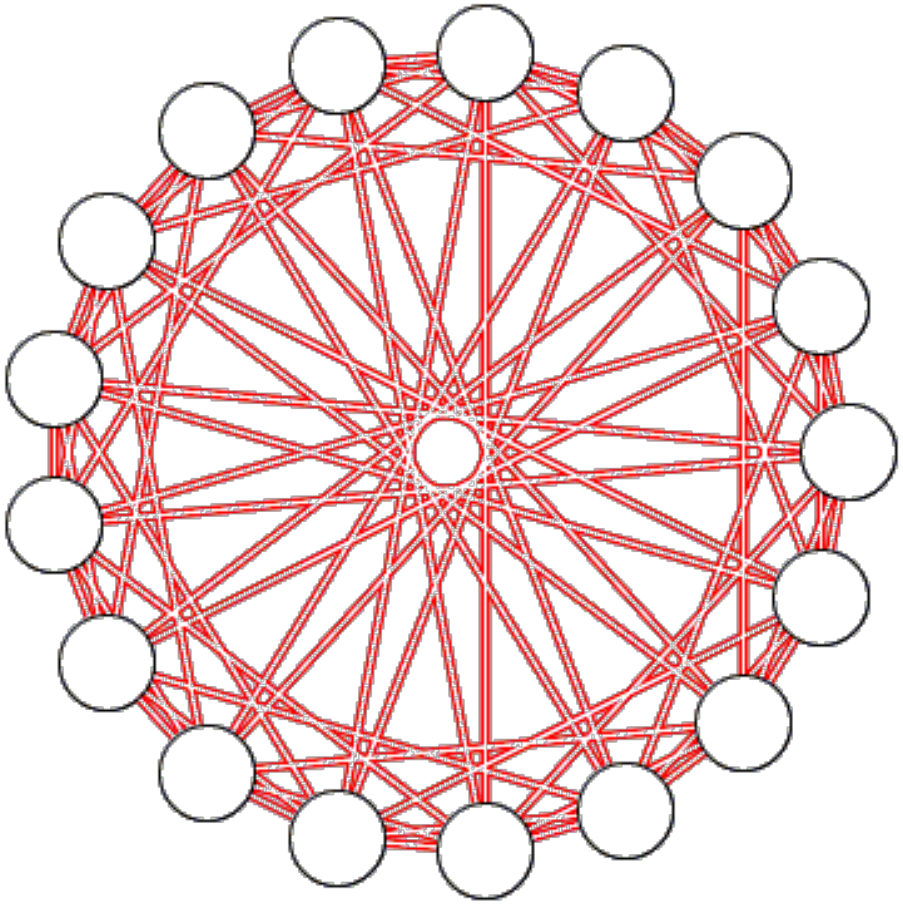
Ramsey's Theorem R(3,3)>5



Ramsey's Theorem R(3,3)=6



Ramsey's Theorem R(4,4)>17



Asymptotic Ramsey Theorem

- Every n-vertex graph must contain a homogenous set of > ½ log n vertices.
- There are n-vertex graphs with no homogenous set of 2 log n vertices. In fact a random G(n, ½) graph has this property.
- The birth of the probabilistic method.

The perspective of graph limits

- We seek an asymptotic theory, i.e., we ask what happens when the number of vertices of the graph $n \to \infty$

- In Math Analysis 101 we learn about limits of sequences of numbers. But how do we develop a limit theory for sequences of graphs?

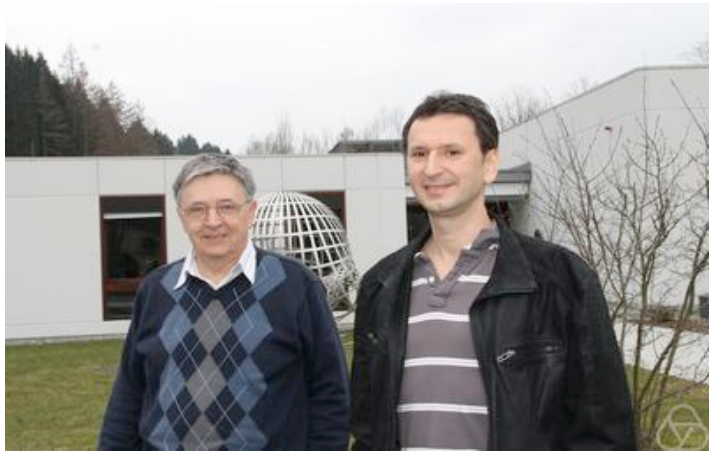
It is well known…

- If you want a limit theory, all you need is a notion of distance (``a metric") d(x,y). You declare that the sequence $x_1, x_2, \cdots$ converges if all distances $d(x_m, x_n)$ are arbitrarily small provided m and n are large enough. (Remember ``Cauchy sequences"?)

- So, how do you measure the distance between two graphs G and H?
- We say that G and H are close if it's possible to chop the vertex sets of both G and of H into N (large integer) equal parts each, so that the following holds: For every $i<j$, the density of the edge set between the i-th and the j-th part in G and in H are nearly equal.
- In words: There is an N-vertex edge-weighted graph that approximates well both G and H.

A key theorem on graph limits

- A theorem of L. Lovasz, B.Szegedy and co. says that a sequence of graphs tends to a limit if and only if the sequence of their local profiles tends to a limit.

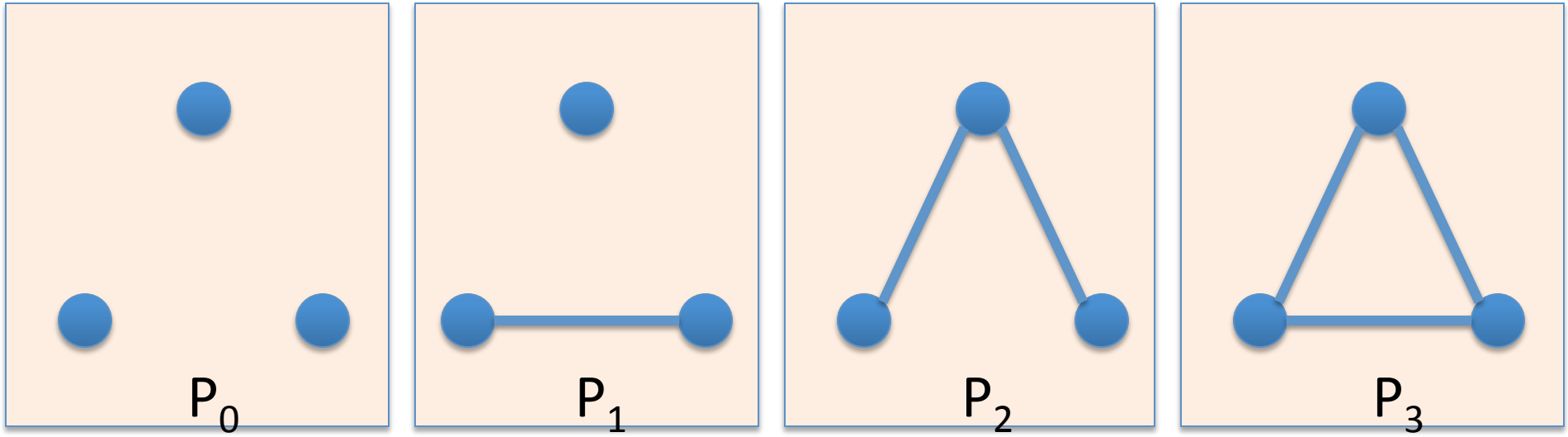


Thus, local profiles may serve as the graph theoretic analog of key statistical parameters such as mean, median, standard deviation etc.

3-profiles

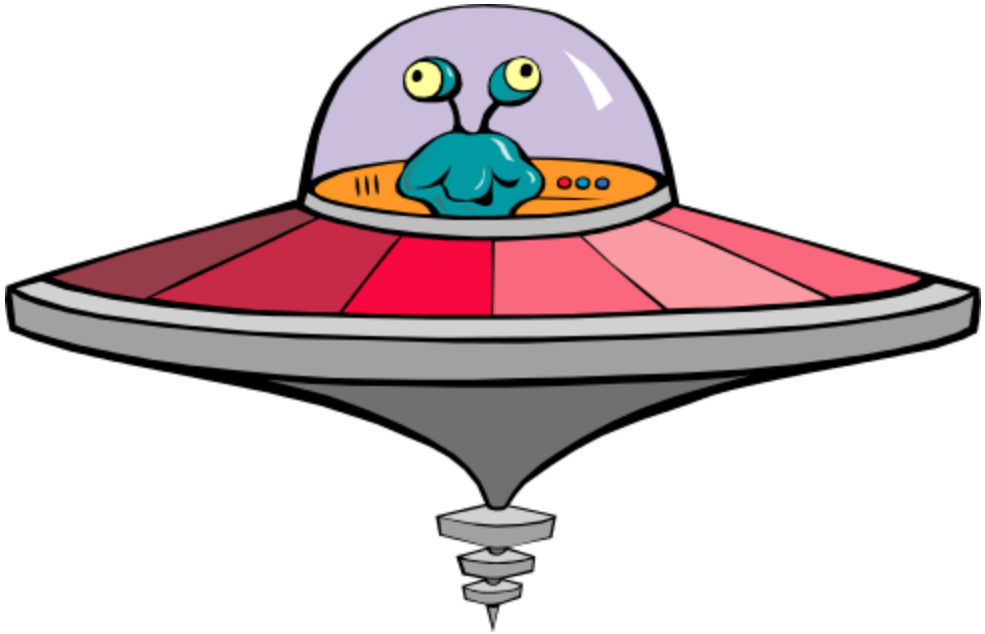
- This is currently the best studied case, but even this is still far from being understood.
- There are four possible 3-vertex graphs that have 0,1,2 and 3 edges.
- We call p_0, p_1, p_2, p_3 the probability of their occurrence respectively.

3-profiles



- Goodman's inequality: $p_0 + p_3 \geq 1/4$
- With Huang, Naves, Peled and Sudakov we proved $min(p_0, p_3) \leq 0.269\ldots$ The bound is tight.

Paul Erdos and the Martians



But even 4-profiles are still completely mysterious to us

- Let us denote by q and r the probability that a set of 4 vertices spans a clique resp. an anticlique. In view of Goodman's inequality the following conjecture is natural, and indeed was made by Erdos:

$$q+r \geq 1/32.$$

Andrew Thomason refuted this

A word on flag algebras

- Recall Turan's theorem for triangles: A graph with density > ½ must contain a triangle.

A word on flag algebras

- Recall Turan's theorem for triangles: A graph with density > ½ must contain a triangle.
- So, e.g., does a graph with density 0.77 necessarily contain <u>many</u> triangles? In words, how small can p_3 be if the density=0.77?

A word on flag algebras

- Recall Turan's theorem for triangles: A graph with density > ½ must contain a triangle.
- So, e.g., does a graph with density 0.77 necessarily contain many triangles? In words, how small can p_3 be if the density=0.77?
- Natural guess: The extreme example for Turan is a bipartite graph with two equal parts. So, try a 5-partite graphs with 4 equal and one smaller parts to achieve right density.

- This was conjectured to be answer, but was open for many years, until proved correct by A. Razborov.

- His main idea: Rather than seek linear inequalities, find quadratic inequalities. Specifically, he has a method to show that certain matrices which capture some of the local structure of graphs are positive semidefinite.
- Computer assisted proofs.

That’s all folks

What does a typical triangle-free graph look like?

- We already saw the simple observation that a bipartite graph contains no triangles.
- For us, these are ``uninteresting" triangle-free graphs.
- What complicates matters is a theorem of Erdos, Kleitman and Rothschild almost all triangle-free graphs are bipartite.
- So how can we sample ``interesting" triangle-free graphs?

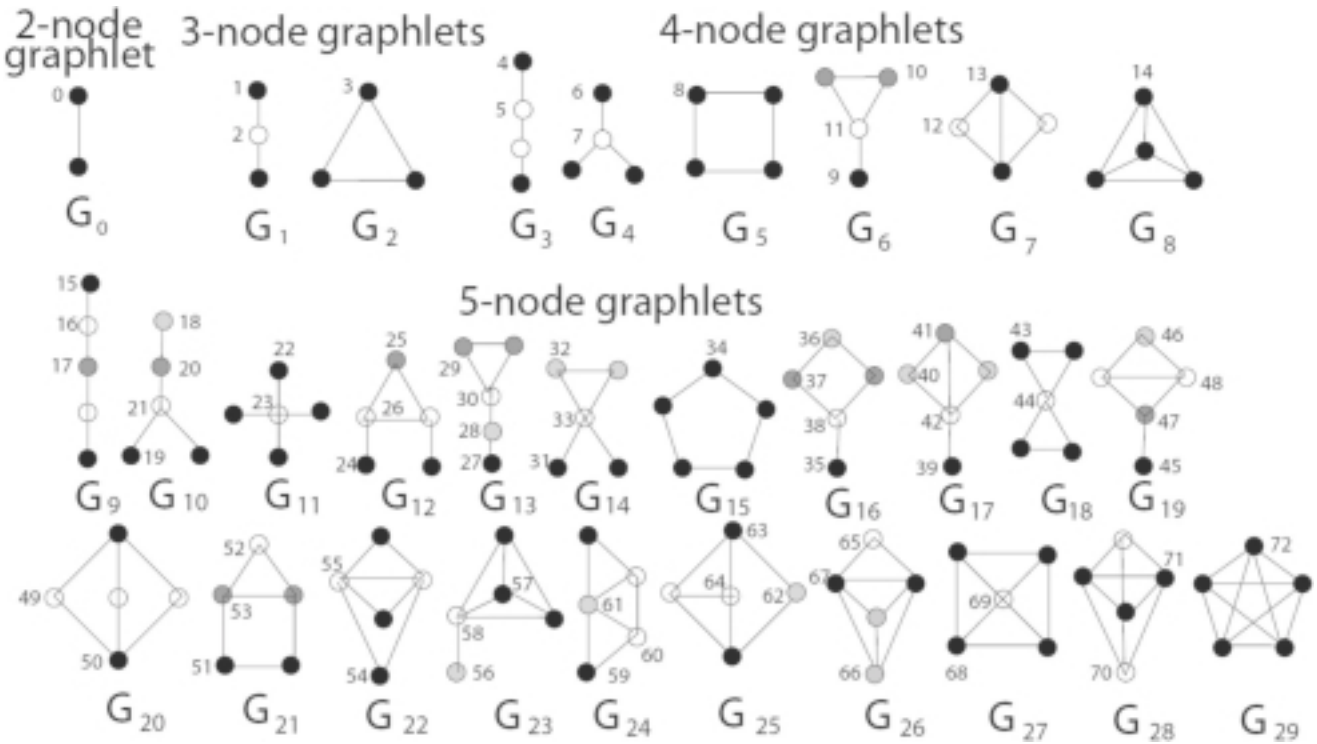
The triangle-free graph process

- Tom Bohman managed to analyze this process, using Wormald's method of differential equations.
- He showed that almost surely this process terminates with a graph that is tight in terms of Ramsey's Theorem.

The triangle-free graph process

- Erdos and Renyi have introduced a close relative of the G(n,p) model, called ``the evolution of random graph".
- This model starts with n vertices and no edges. Sequentially at each step a new random edge is added.
- The triangle-free process does the same, except that if a prospective new edges closes a triangle, it's discarded.

Counting..


2-node graphlet
3-node graphlets
4-node graphlets
5-node graphlets
G0
G1
G2
G3
G4
G5
G6
G7
G8
G9
G10
G11
G12
G13
G14
G15
G16
G17
G18
G19
G20
G21
G22
G23
G24
G25
G26
G27
G28
G29