Testing Cluster Structure of Graphs

Artur Czumaj

DIMAP and Department of Computer Science

University of Warwick

Joint work with Pan Peng and Christian Sohler (TU Dortmund)

Dealing with “BigData” in Graphs

- We want to process graphs quickly
  - Detect basic properties
  - Analyze their structure

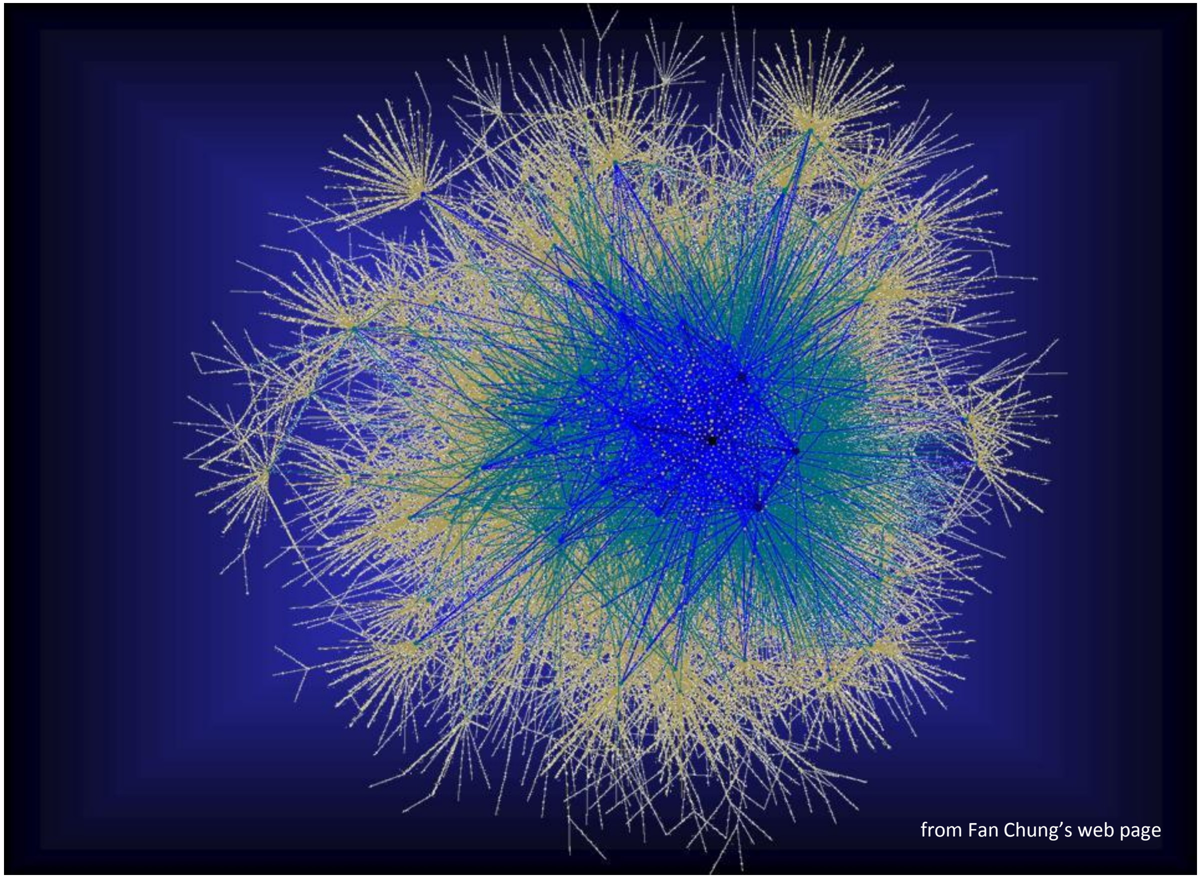
- For large graphs, by “quickly” we often would mean: in time *constant* or *sublinear* in the size of the graph

Dealing with "BigData" in Graphs

One approach:

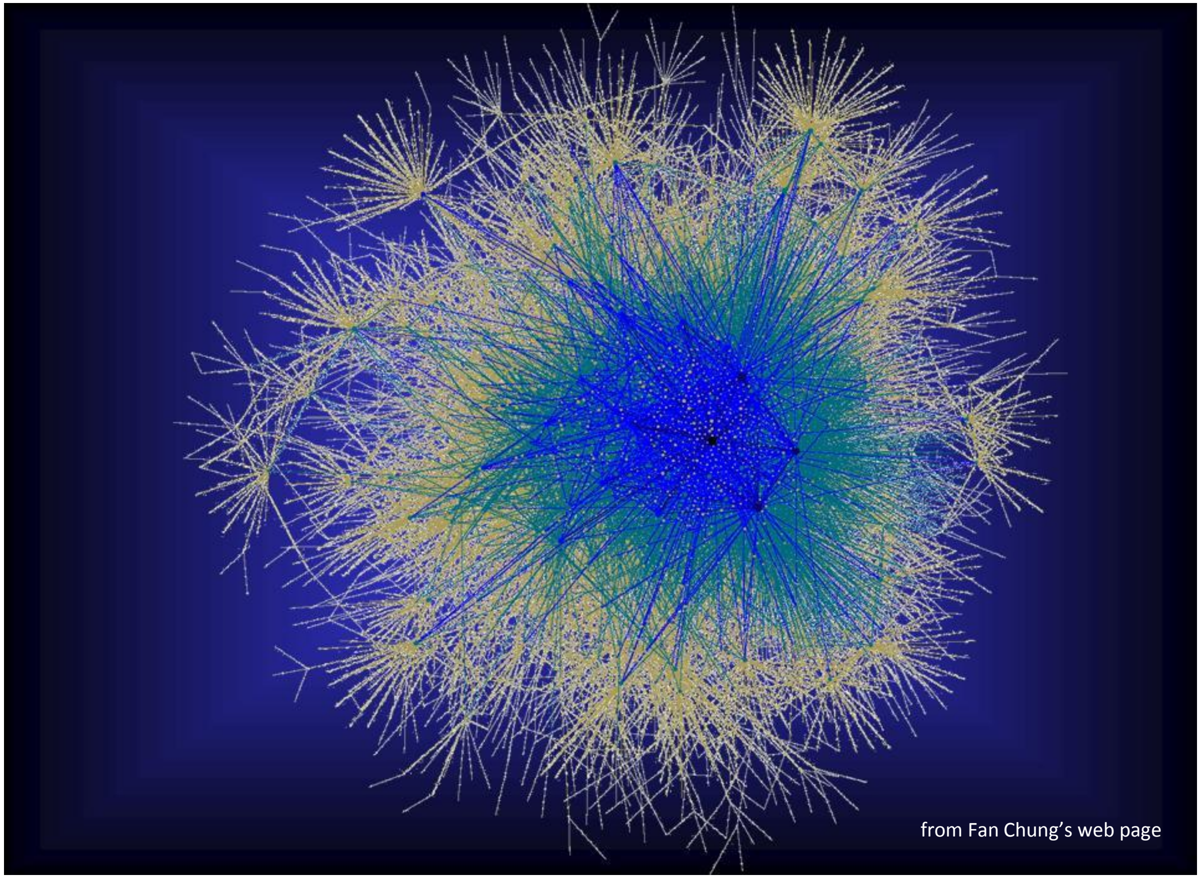
- How to test basic properties of graphs
  in the framework of **property testing**

Fast Testing of Graph Properties



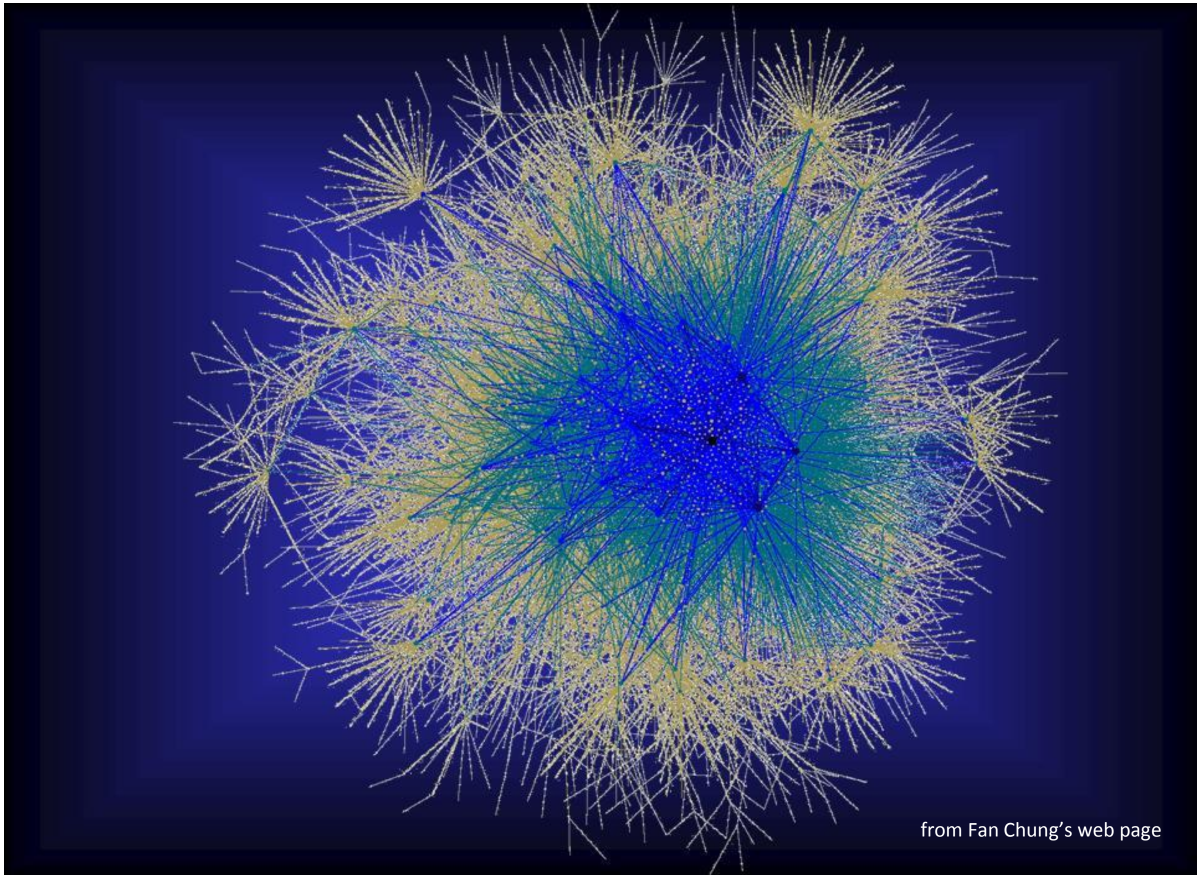
- Does this graph have a clique of size 11?
- Does it have a given H as its subgraph?
- Is this graph planar?
- Is it bipartite?
- Is it k-colorable?
- Does it have good expansion?
- Does it have good clustering?

Clustering in graphs



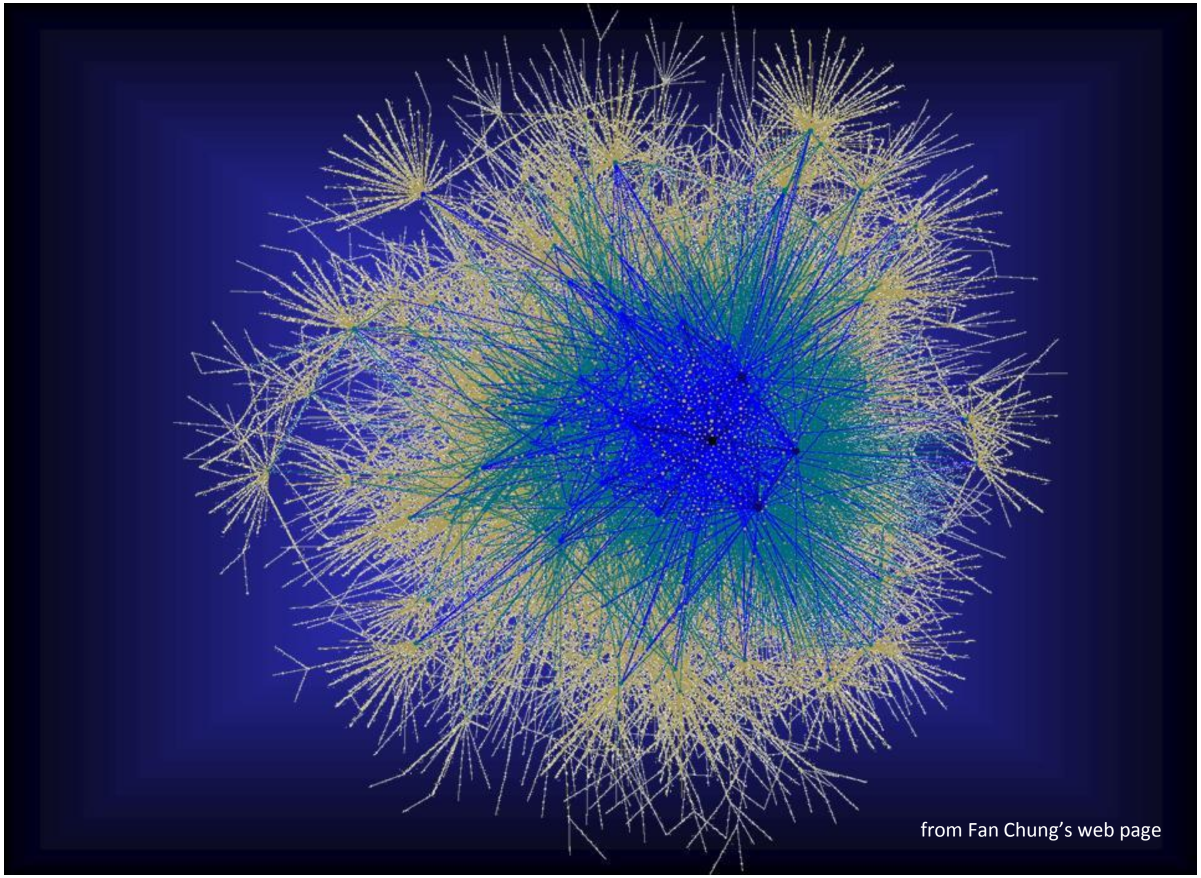
- What is a good clustering?

Clustering in graphs


from Fan Chung's web page

- Same cluster: points are well-connected
- Different cluster: points are poorly connected

Clustering in graphs



- Same cluster: points have **high conductance**
- Different cluster: points are **separated by a cut**

k-clustering

Graph G is ***k*-clusterable** if vertices of G can be partitioned into at most k sets $V_1, V_2, \ldots$ such that for each i:

- each $G[V_i]$ has large conductance
- each set V_i has low outer-conductance (small cut)

Conductance

G has maximum degree $\leq d$

For every $S \subseteq V$, conductance of S is defined as:

$$\phi_G(S) = \frac{e(S, V \setminus S)}{d\ |S|}$$

Conductance of G:

$$\phi_G = \min_{S \subseteq V, |S| \leq |V|/2} \phi_G(S)$$

(minimum conductance of any possible subset of size at most $|V|/2$)

$(k, \phi_{in}, \phi_{out})$-clustering

G is **$(k, \phi_{in}, \phi_{out})$-clusterable** if vertices of G can be partitioned into $V_1, V_2, \dots, V_s$ with $s \leq k$ such that for each i:

- each $G[V_i]$ has large conductance

  $\phi(G[V_i]) \geq \phi_{in}$

- each set V_i has low outer-conductance (small cut)

  $\phi_G(V_i) \leq \phi_{out}$

$(k, \phi_{in}, \phi_{out})$-clustering

Notion of $(k, \phi_{in}, \phi_{out})$-clusterable graphs has been around informally for a while;

formally introduced by Oveis Gharan and Trevisan 2014

Our goal:

- we want to determine if G is $(k, \phi_{in}, \phi_{out})$-clusterable
- really fast
  - Recognize cluster structure in sublinear time using random sampling

Framework of property testing

- We cannot quickly give 100% precise answer
- We need to approximate

- Distinguish graphs that have specific property from those that are far from having the property

Property Testing definition

- Given input G
- If G has the property ⇨ tester passes
- If G is ε-far from any string that has the property ⇨ tester fails
- error probability < 1/3

Notion of **ε-far** : DISTANCE to the Property
One needs to change ε fraction of the input to obtain an object satisfying the property

Typically we think about $\boldsymbol{\varepsilon}$
as on a small constant, say, $\boldsymbol{\varepsilon = 0.1}$

Framework

- Goal:

  Distinguish between the case when
  - graph G has property P and
  - G is far from having property P
    - *one has to change at least εdn edges of G to obtain a graph with property P*

We will consider graph of degree bounded by d

Goal

Design a sublinear-time algorithm that will distinguish between two cases:

- $(k, \phi_{in}, \phi_{out})$-clusterable graphs
- graphs that are ε-far from being $(k, \phi_{in}^*, \phi_{out}^*)$-clusterable (with ϕ_{in}^* as close to ϕ_{in} as possible)

Basic case $k = 1$: Testing expansion

$(k, \phi_{in}, \phi_{out})$-clusterable graphs for $k = 1$: expanders

- For graphs of bounded degree, we can distinguish expanders from graphs that are "far" even from poor expanders in $O^*(\sqrt{n})$ time

  [C, Sohler '07, Kale, Seshadhri'07, Nachmias, Shapira'08]
- $\Omega(\sqrt{n})$ time is needed

  [Goldreich, Ron'02]

Basic case $k = 1$: Testing expansion

- For graphs of bounded degree, we can distinguish expanders from graphs that are "far" even from poor expanders in $O^*(\sqrt{n})$ time
  [C, Sohler '07, Kale, Seshadhri'07, Nachmias, Shapira'08]

- We are using basic properties of expanders: random walk of logarithmic length will mix (= will reach a random vertex)
- Similar to testing uniformity of a distribution

Testing expansion

Choose $O(1/\varepsilon)$ nodes at random
For each chosen node run $O(\sqrt{n})$ random walks of length $O(\log n)$
    Count the number of collisions at the end-nodes
    If the number of collisions is too large then **Reject**
Accept

Idea:

- If G is an expander then end-nodes are random nodes

  ⇨ we can estimate number of collisions well

- If G is far from expander then we will have many more collisions
  - (requires non-trivial arguments)

Basic case $k = 1$: Testing expansion

- For graphs of bounded degree, we can distinguish expanders from graphs that are "far" even from poor expanders in $O^*(\sqrt{n})$ time

  [C, Sohler '07, Kale, Seshadhri'07, Nachmias, Shapira'08]

- We can distinguish between a graph G that is an λ-expander and any graph that is ε-far from any $c\lambda^2/d$-expander in time $O^*(n^{0.5+\delta} f(\epsilon))$

Testing expansion and clustering

Can we apply similar approach to test $(k, \phi_{in}, \phi_{out})$-clusterability for $k > 1$?

- We don't know which vertex sets form expanders
- We don't know sizes of subgraphs-expanders
  - If we knew, we could try to test distributions of endpoints of random walks …
- How to test small cuts?
- We don't know how to test distributions in $o(n^{2/3})$ time!
  - We know this only for uniform distributions – but since we don't know which vertices are in each set, we won't get it …

Testing expansion and clustering

Can we apply similar approach to test $(k, \phi_{in}, \phi_{out})$-clusterability for $k > 1$?

Still, we will following the following key intuitions:

- Randomly sample a constant number of points S
- Points in S will define a "skeleton" of $\leq k$ clusters
- If two points will have same distribution of end-points of random walks of logarithmic length ➔ are in same cluster
- If two points are separated by a cut then they will have different distribution of end-points of random walks

Testing $(k, \phi_{in}, \phi_{out})$-clustering

- We would like to have the following algorithm:

Sample set S of s random vertices
For any $v \in S$
- D_v =distributions of endpoints of random walk of length ℓ starting at v

For each pair $u, v \in S$:
- if distributions D_u, D_v are close then
  add edge (u, v) to "cluster graph" H on vertex set S

If H is a union of at most k cliques then **Accept**
Else **Reject**

- Too slow (testing if distribution are closed needs $\Omega(n^{2/3})$ time)
- How to analyze it ?
- We understand random walks in expanders, but we need to understand them also on the rest of the graph

Testing $(k, \phi_{in}, \phi_{out})$-clustering

Sample set S of s random vertices
For any $v \in S$

- $F_v =$ multiset of endpoints of r random walks of length ℓ starting at v
- $Z_v =$ number of pairwise self-collisions in F_v

If there is $v \in S$ with $Z_v > \sigma$ then **abort** and **Reject**
For each pair $u, v \in S$:

- If ℓ_2-distribution-test (F_u, F_v) accepts that distributions of F_u and F_v are close then add edge (u, v) to "cluster graph" H

If H is a union of at most k cliques then **Accept**
Else **Reject**

$s = O(k \ln(k+1)\, \varepsilon^{-2})$

$r = O(k^2 (\ln k/\varepsilon)^{5/2} \sqrt{n} \varepsilon^{-3})$

$\ell = O(k^4 \log n\, \phi_{in}^{-2})$

$\sigma = O(k^6 \ln(k/\varepsilon)^{6}\, \varepsilon^{-8})$

Key theorem

- The algorithm accepts every $(k, \phi_{in}, \phi_{out})$-clusterable graph (of maximum degree $\leq d$) with probability $\geq 2/3$.

- The algorithm rejects every graph G (of maximum degree $\leq d$) that is ε-far from being $(k, \phi_{in}^*, \phi_{out}^*)$-clusterable with probability $\geq 2/3$, assuming that $\phi_{in}^* \leq c\, \phi_{in}^2 \varepsilon^4 / \log n$.

- Running time is $\sqrt{n}\, \phi_{in}^{-2} (k\varepsilon^{-1} \log n)^{O(1)}$

Key properties (completeness)

- $\boldsymbol{p}_u^t$- vertex distribution of a random walk of length t starting at u

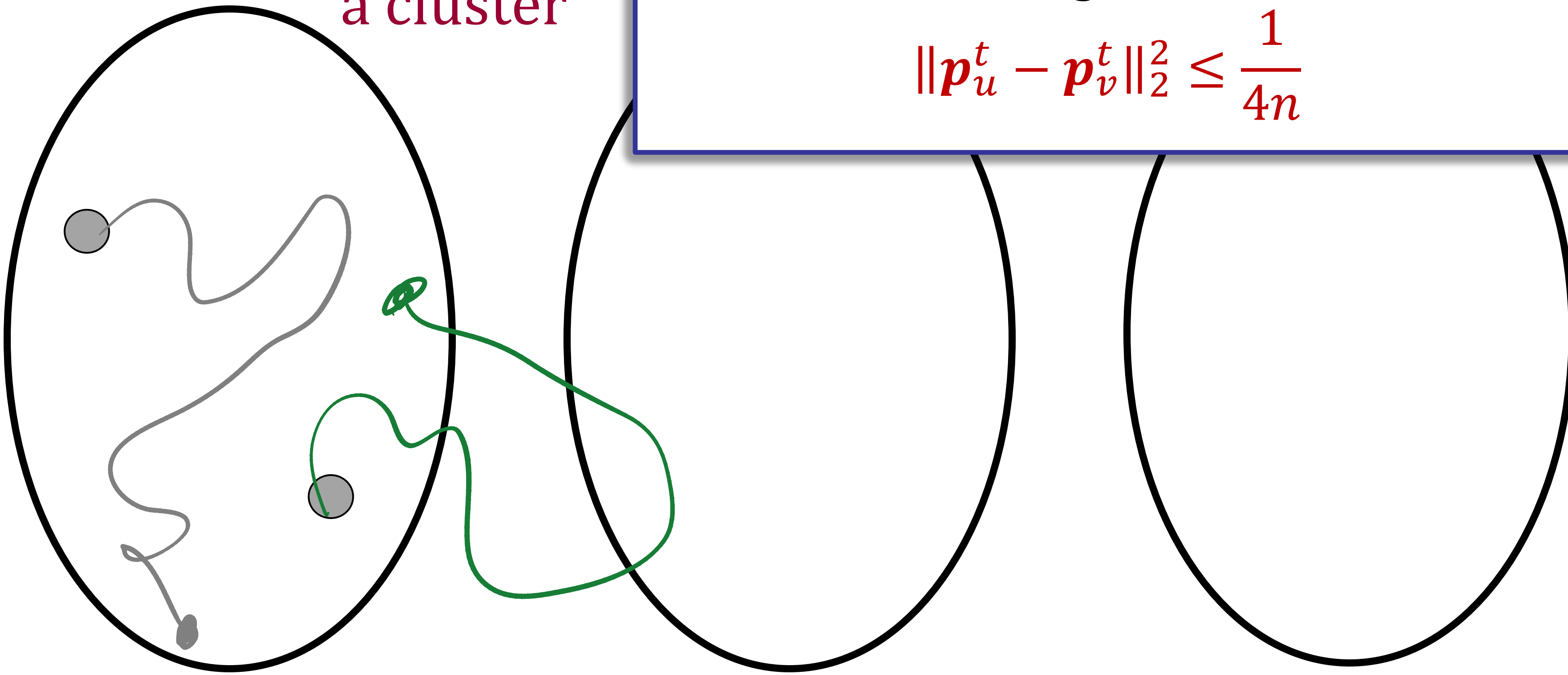
Key properties (completeness)

- Convergence within a cluster

How will the distribution of endpoints differ (for clusterable graphs)?

For most of starting vertices u, v:

$$\|\boldsymbol{p}_u^t - \boldsymbol{p}_v^t\|_2^2 \leq \frac{1}{4n}$$



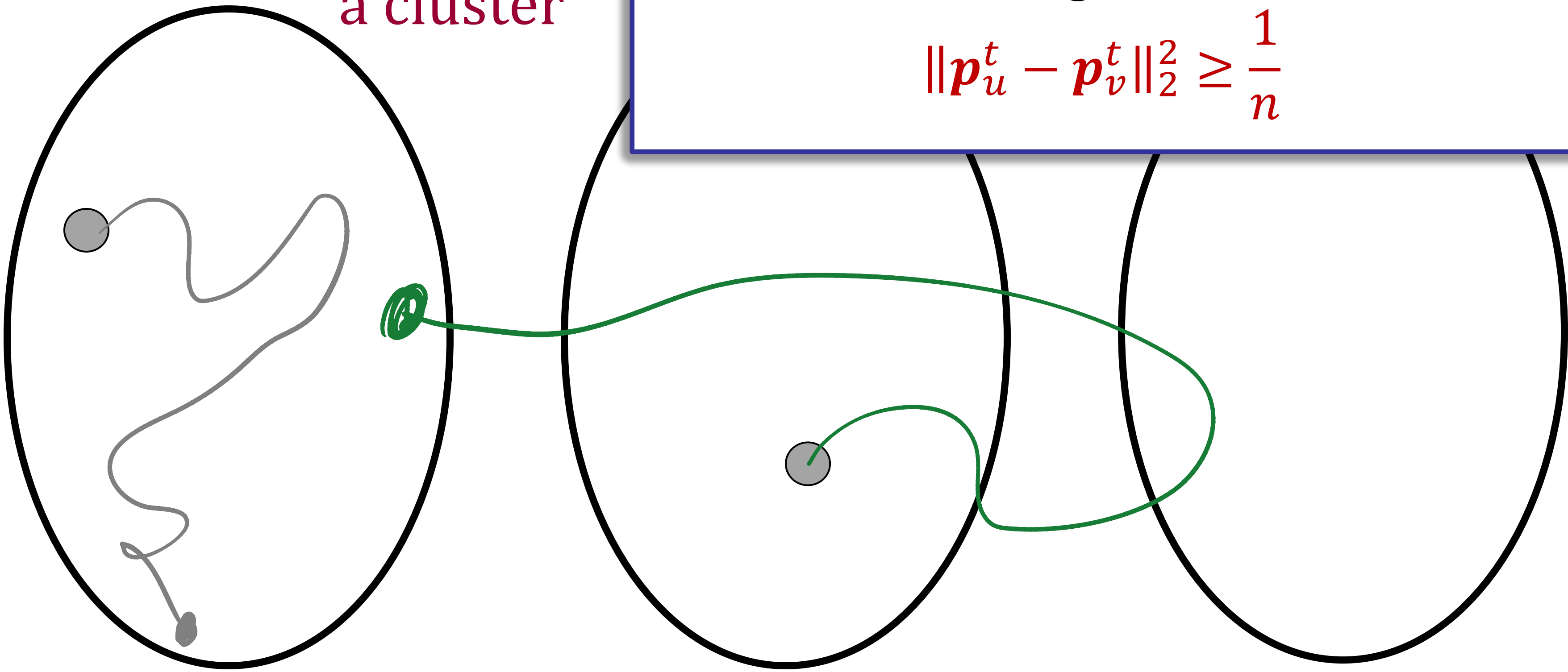
Key properties (completeness)

- Convergence outside a cluster

How will the distribution of endpoints differ?

For most of starting vertices u, v:

$$\|\boldsymbol{p}_u^t - \boldsymbol{p}_v^t\|_2^2 \geq \frac{1}{n}$$



Key properties (completeness)

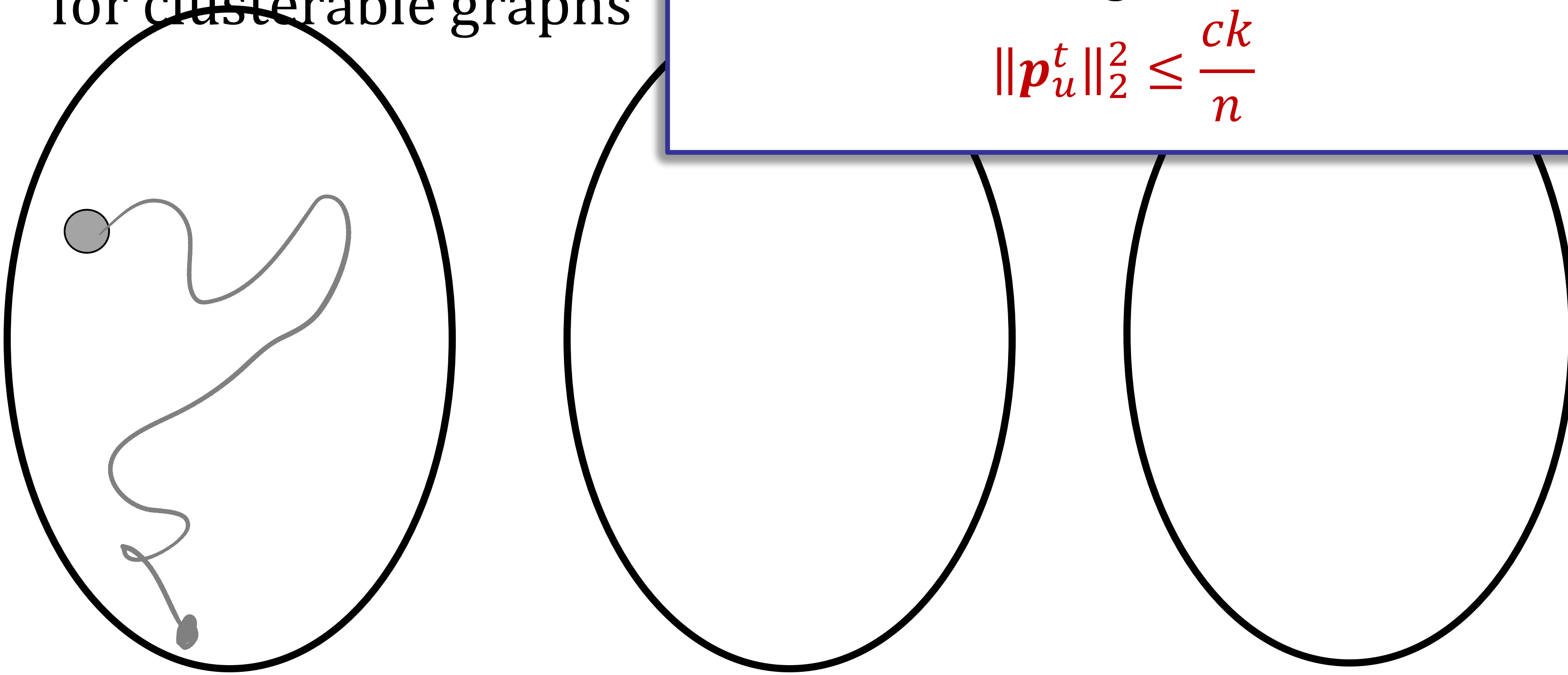
- Bound for distribution

  for clusterable graphs

Can the distribution vector of endpoints have big values?

For most of starting vertices u:

$$\|\boldsymbol{p}_u^t\|_2^2 \le \frac{ck}{n}$$



Key properties (completeness)

- $\boldsymbol{p}_u^t$- vertex distribution of a random walk of length t starting at u
- If G is $(k, \phi_{in}, \phi_{out})$-clusterable then for any $C \subseteq V$ with $|C| \geq \beta|V|$ and $\phi(G[C]) \geq \phi_{in}$ there is $C^* \subseteq C$ with $|C^*| \geq (1-\alpha)|C|$ such that for large enough t and $\phi_{out} \leq c\phi_{in}/\log^2 n$, for any $u, v \in C^*$:
$$\|\boldsymbol{p}_u^t - \boldsymbol{p}_v^t\|_2^2 \leq \frac{1}{4n}$$
- For “short enough” t, for any disjoint sets $S, T \subseteq V$ with $\phi_G(S), \phi_G(T) \leq \psi$, there exist $S^* \subseteq S, T^* \subseteq T$, $|S^*| \geq (1-\alpha)|S|$, $|T^*| \geq (1-\alpha)|T|$ such that for any $u \in S^*, v \in T^*$
$$\|\boldsymbol{p}_u^t - \boldsymbol{p}_v^t\|_2^2 \geq \frac{1}{n}$$
- If G is $(k, \phi_{in}, \phi_{out})$-clusterable then there is $V^* \subseteq V$ with $|V^*| \geq (1-\alpha)|V|$ such that for large enough t, for any $u \in V^*$:
$$\|\boldsymbol{p}_u^t\|_2^2 \leq \frac{2k}{\alpha n}$$

Key properties (completeness)

With these properties:

- If G is clusterable then the cluster graph H will consist of at most k disjoint subgraphs, each forming a clique

Key properties (soudness)

- If G is ε-far from $(k, \phi_{in}^*, \phi_{out}^*)$-clusterable with $\phi_{in}^* \le c\ \varepsilon$ then there are $k+1$ disjoint subsets $V_1, V_2, \dots, V_{k+1}$ of V such that for each i:

$$|V_i| \ge c\varepsilon^2 |V|/k \text{ and } \phi_{G(V_i)} \le c\phi_{in}^* \varepsilon^{-2}$$

With this property, if G is ε-far from clusterable then the cluster graph H will have more than k components

Key theorem

- The algorithm accepts every $(k, \phi_{in}, \phi_{out})$-clusterable graph (of maximum degree $\leq d$) with probability $\geq 2/3$.

- The algorithm rejects every graph G (of maximum degree $\leq d$) that is ε-far from being $(k, \phi_{in}^*, \phi_{out}^*)$-clusterable with probability $\geq 2/3$, assuming that $\phi_{in}^* \leq c\, \phi_{in}^2 \varepsilon^4 / \log n$.

- Running time is $\sqrt{n}\, \phi_{in}^{-2} (k\varepsilon^{-1} \log n)^{O(1)}$

Extensions

Since k-clustering is related to some properties of k smallest eigenvalues of the relevant Laplacian matrix:

- We can recognize graphs with a (large enough) gap between the k^{th} and $k+1^{st}$ smallest eigenvalue.

Conclusions

Clustering (or clusterability) can be tested fast

- by comparing distributions of random walks
- drawing conclusions from the distributions

Tools:

- Random sampling
- Random walks
- Spectral analysis