The Power of a Few Local Samples for Predicting Epidemics

Yeganeh Alimohamamdi

Stanford University

Joint work with Amin Saberi, Christian Borgs, and Remco van der Hofstad

Motivation and outline

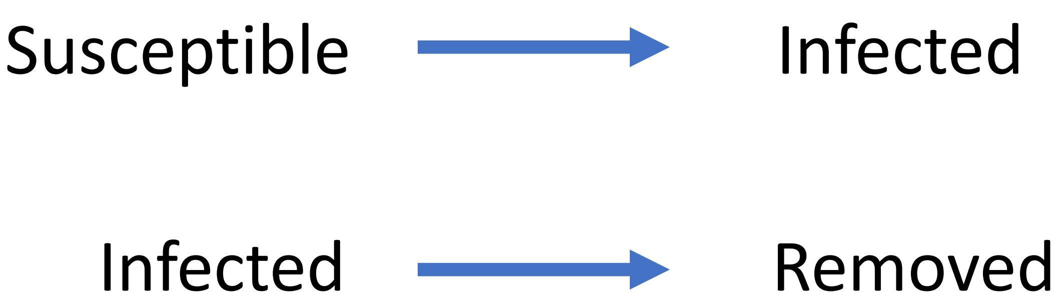
- Predicting epidemics: important for risk assessment, evaluating the effectiveness of countermeasures, and seeding

- This talk: simple algorithms for estimating the size & likelihood of outbreaks, and their time evolution

- Main technical ingredient: graph limits, a generalization for most existing network models

The epidemic model

Susceptible-**I**nfected-**R**ecovered (SIR)



Susceptible → Infected

at rate β(no of infected contacts)

Infected → Removed

after some time T e.g.
Poisson with rate γ (contact process)
constant

[Ross and Hudson 1916, Kermack, McKendrick 1927]

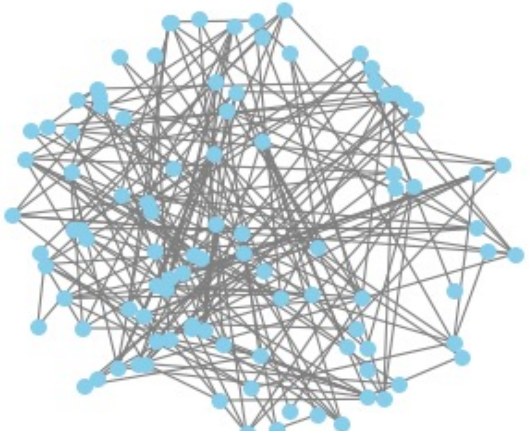
Predicting the spreading of an infection

- **Mean-Field Models** (Curie-Weiss, cf. Keeling, Rohani ’07]
  - Average over the whole network
  - Does not capture stochasticity of the process

- **Random graph Models** (cf. Bollobas 2011, Durrett 2005)
  - Stochastic processes on random graphs
  - Relies on the estimation of network parameters

Epidemics on random networks

- **Erdos-Renyi graphs** aka $G(n,p)$
  each pair is connected independently with probability p



Epidemics on random networks

- **Erdos-Renyi graphs** aka $G(n, p)$

outbreak if $$\underbrace{(n-1)p}_{\text{expected degree}} \underbrace{\frac{\beta}{\gamma}}_{\text{transmission probability}} \geq 1$$

Epidemics on random networks

- **Erdos-Renyi graphs** aka $G(n, p)$

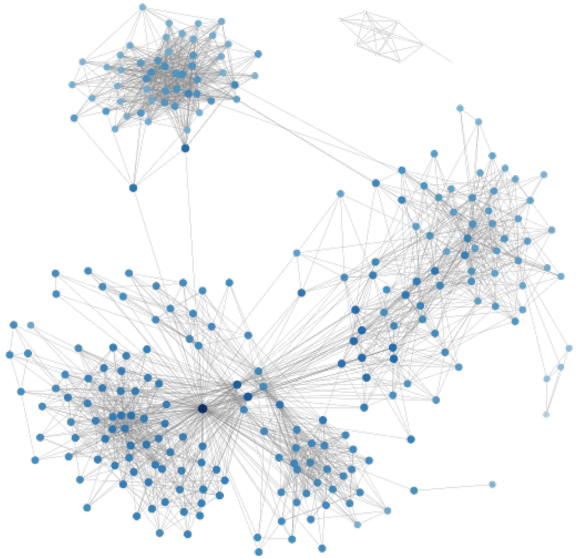
outbreak if $(n-1)p\ \frac{\beta}{\gamma} \geq\ 1$

- **Power-law random graphs:** configuration model [Molloy-Reed, Barabasi, Watts '11], preferential attachment [Bollobás-Riordan '03]

outbreak if $\frac{\overline{d^2-d}}{\bar{d}} \frac{\beta}{\gamma} \geq 1$

Second moment

expected degree



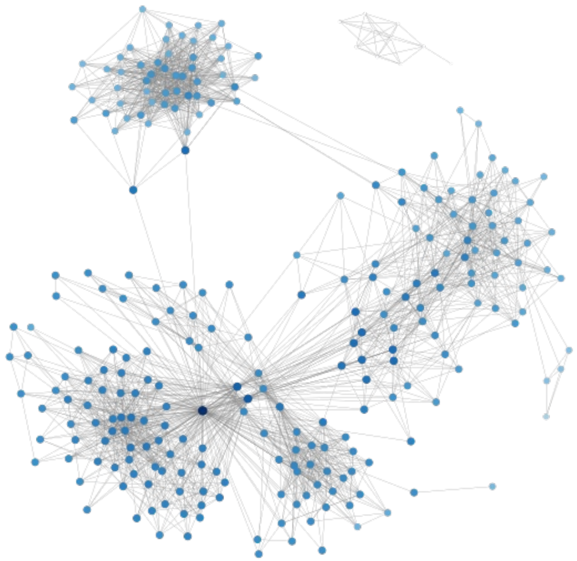
Epidemics on random networks

- **Erdos-Renyi graphs** aka $G(n, p)$

outbreak if $(n-1)p\,\frac{\beta}{\gamma} \geq 1$

- **Power-law random graphs:** configuration model [Molloy-Reed, Barabasi, Watts ‘11], preferential attachment [Bollobás-Riordan ‘03]

outbreak if $\frac{\overline{d^2-d}}{\bar{d}}\frac{\beta}{\gamma} \geq 1$



Many others e.g. stochastic block model, household models…

Predicting the spreading of an infection

- **Mean-Field Models** (Curie-Weiss, cf. Keeling, Rohani ’07]
  - Average over the whole network
  - Does not capture stochasticity of the process

- **Random graph Models** (cf. Bollobas 2011, Durrett 2005)
  - Stochastic processes on random graphs
  - Relies on the estimation of network parameters

- **Stochastic simulation on the network**
  - Requires complete network access (expensive, privacy concerns)
  - hard to determine robustness to network mis-specification

Summary

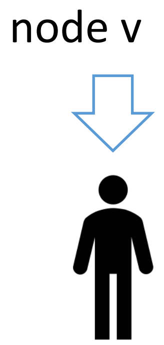
Stylized models (using mean-field calculations or random graphs) are great for gaining insights but fall short for prediction tasks. On the other hand, simulations on the whole network are often costly or even impossible.

This talk in a nutshell:

Under general assumptions, local information about a small sample of nodes is sufficient for estimating the final size and the time evolution of the epidemics.

Local estimator

How to estimate the probability of states (S,I, R) of node v at time t?



<u>Naive Method</u>

Input: k, and time t

- Find k-hop neighborhood of v
- Simulate infection until time t
- Return state of v

Local estimator

How to estimate the probability of states (S,I, R) of node v at time t?

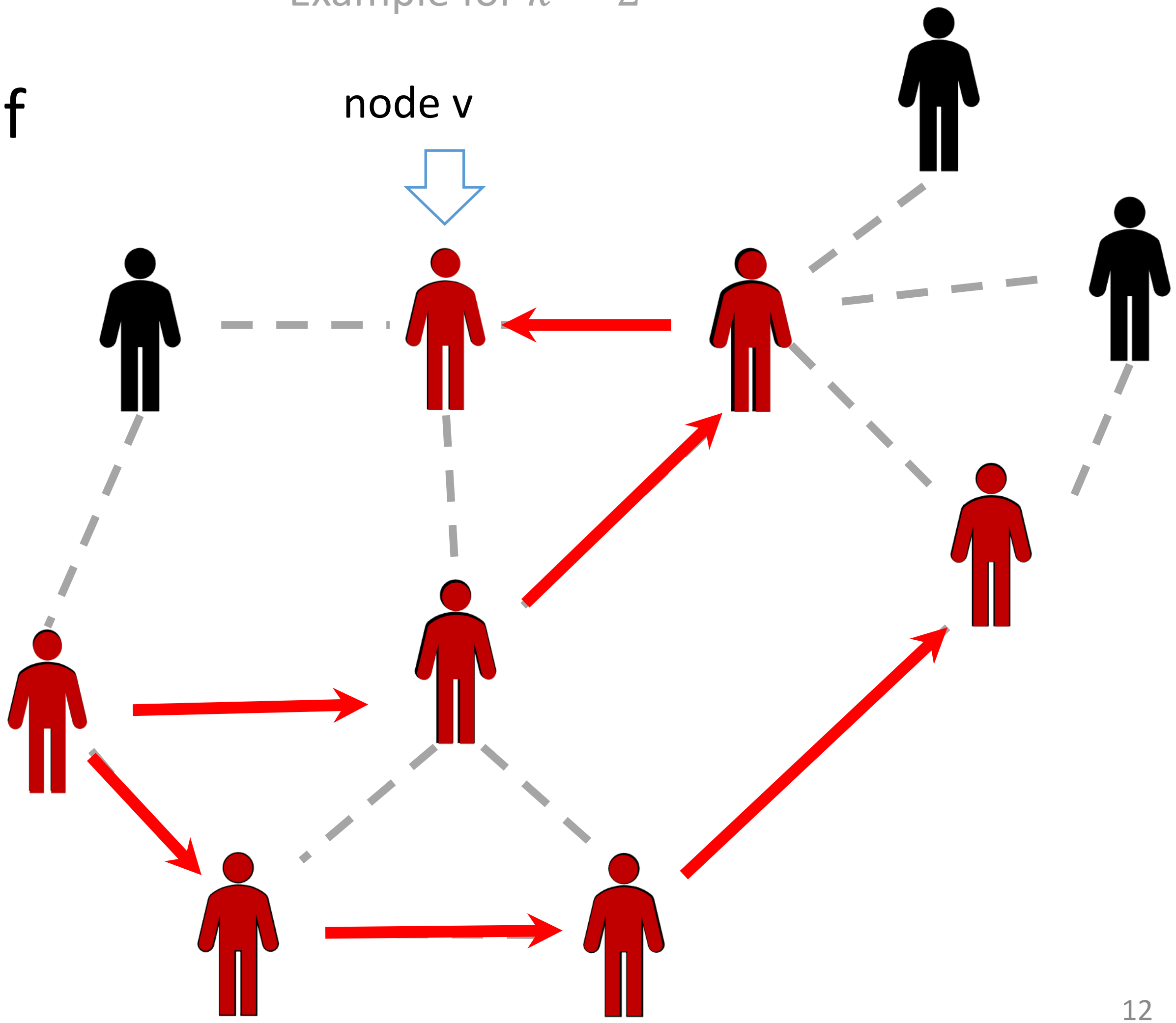
<u>Naive Method</u>

Input: k, and time t

- ○ Find k-hop neighborhood of v
- ○ Simulate infection until time t
- ○ Return state of v

Exponential growth of nodes explored (in k)

Example for $k = 2$



A better implementation: backward process

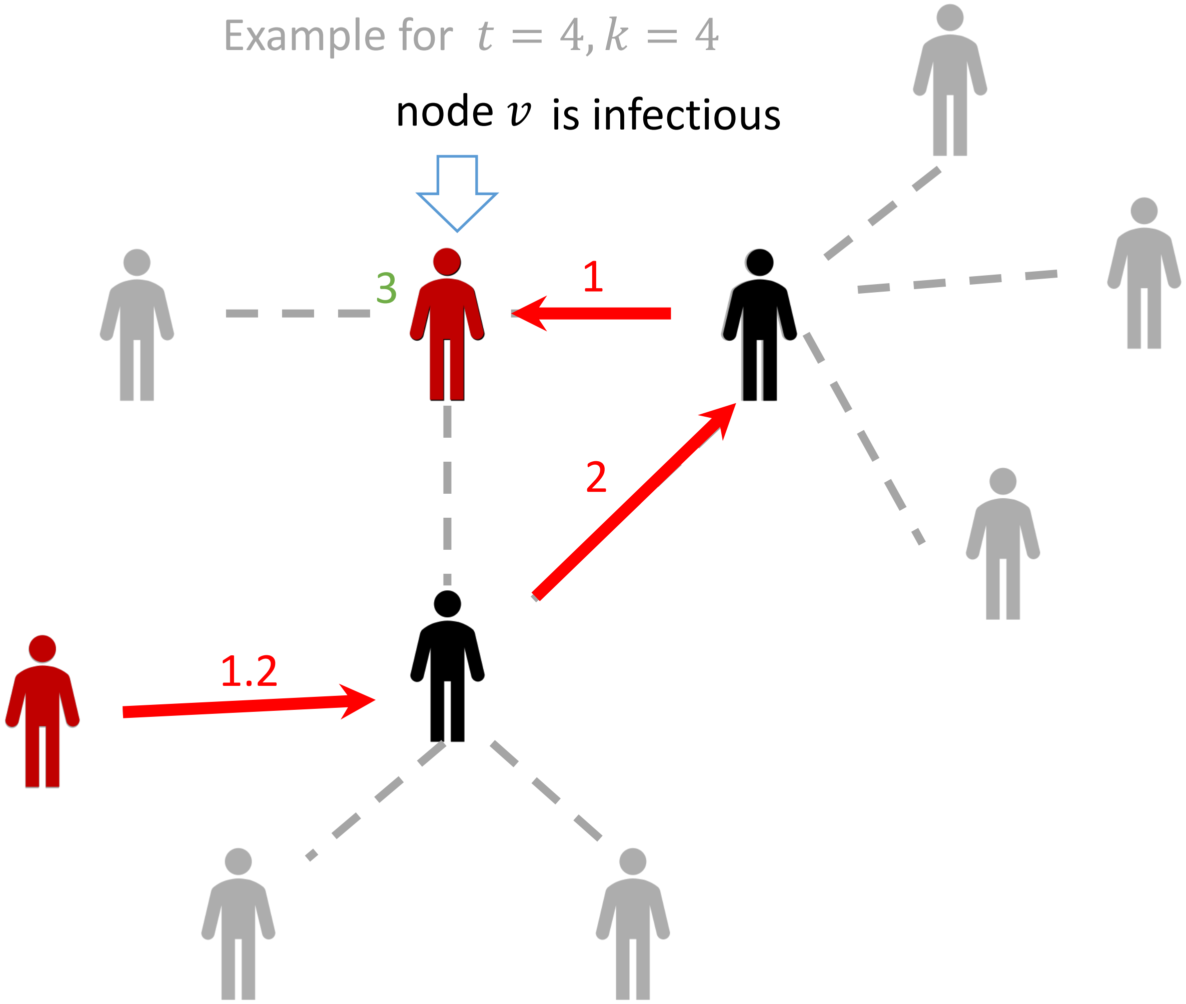
Simulates the state of a uniformly random node v at time t going backwards in time

<u>Timed Backward Process</u>

Input: k, and time t

- ○ Simulate infection backward until see k people or reach time t
- ○ Return state of v

Example for $t = 4, k = 4$



A better implementation: backward process

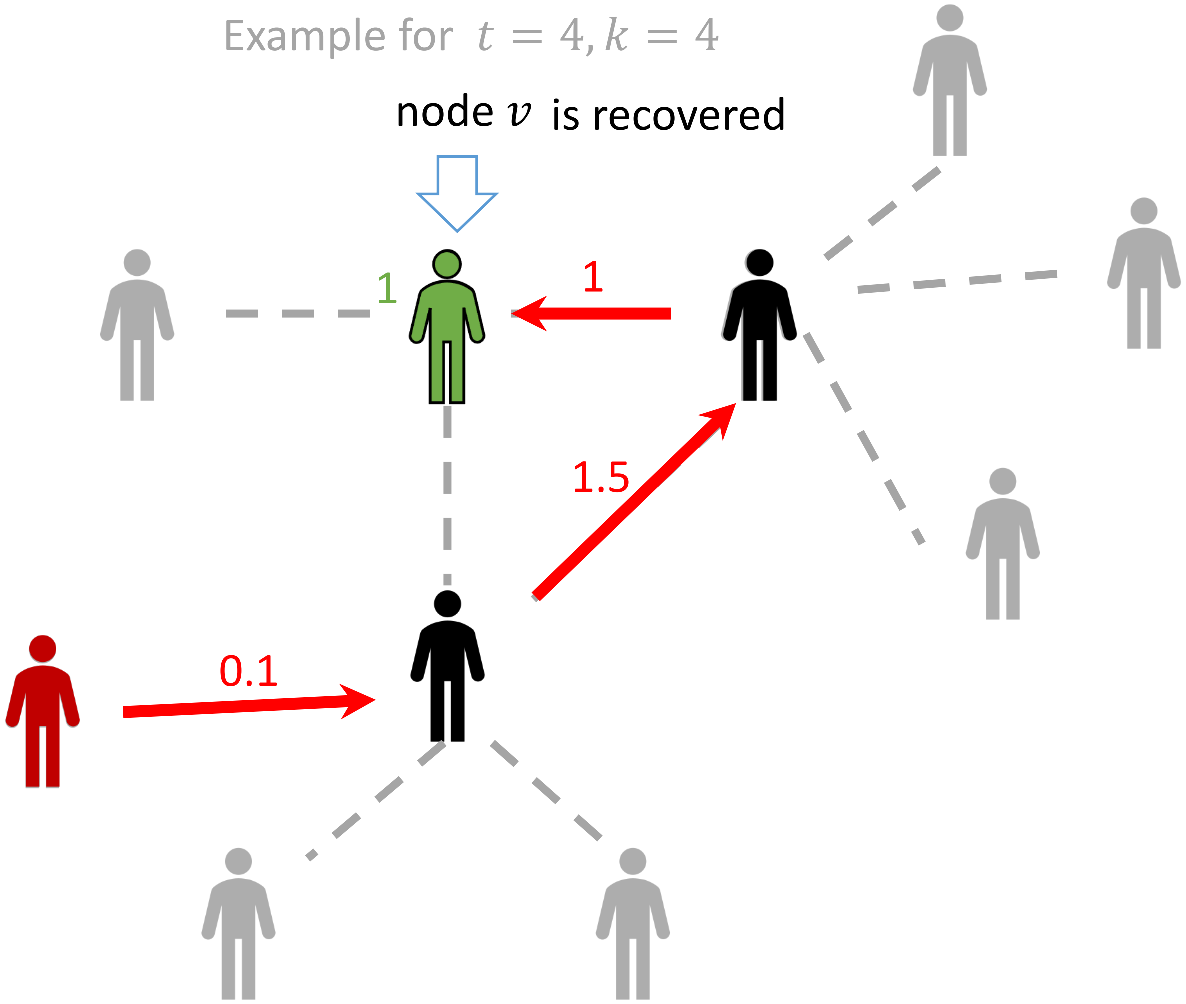
Simulates the state of a uniformly random node v at time t going backwards in time

<u>Timed Backward Process</u>

Input: k, and time t

- ○ Simulate infection backward until see k people or reach time t
- ○ Return state of v

Example for $t = 4, k = 4$



A better implementation: backward process

Simulates the state of a uniformly random node v at time t going backwards in time

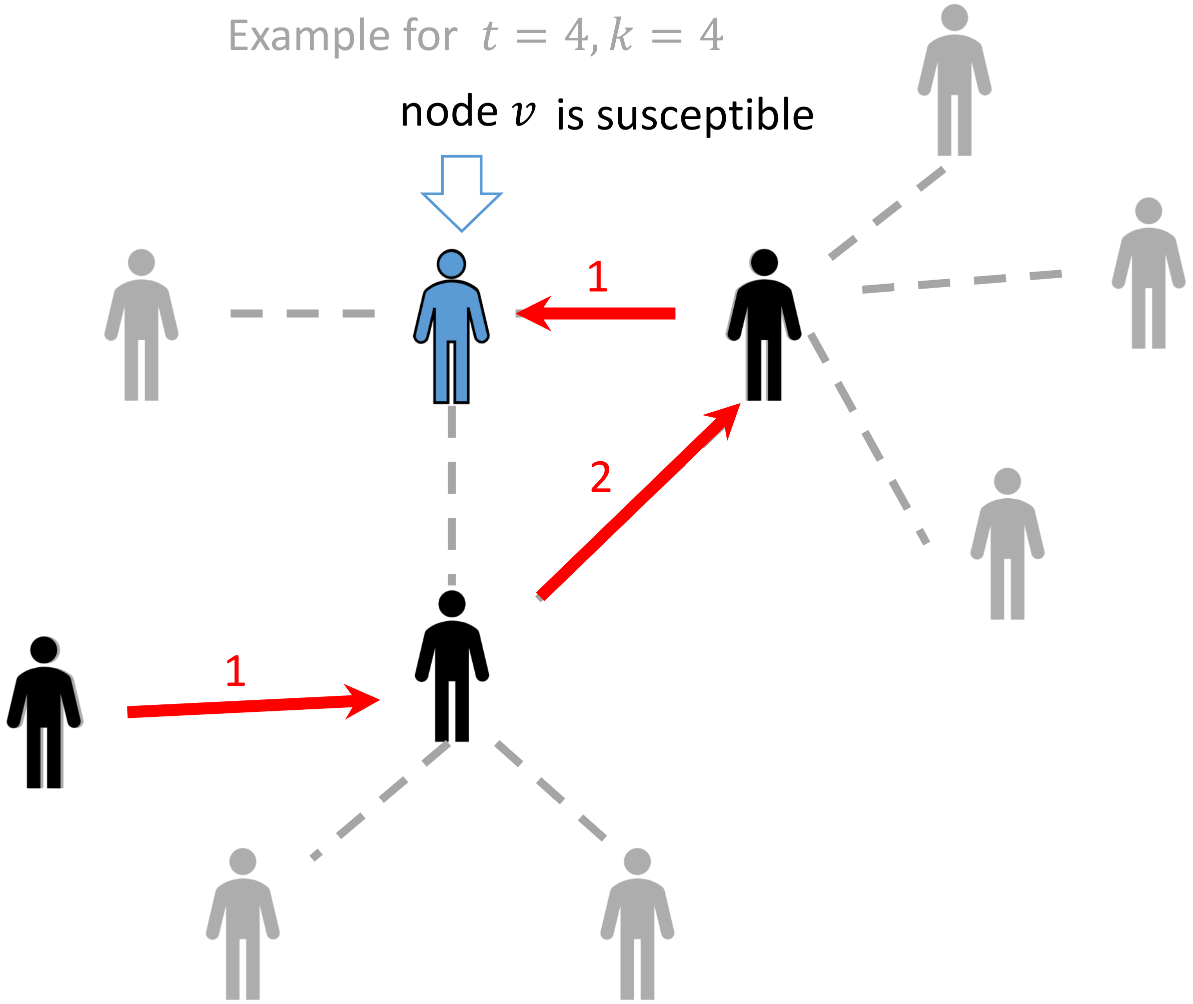
<u>Timed Backward Process</u>

Input: k, and time t

- ○ Simulate infection backward until see k people or reach time t
- ○ Return state of v

Estimator: average over queries

Example for $t = 4, k = 4$



Key features of the local estimators

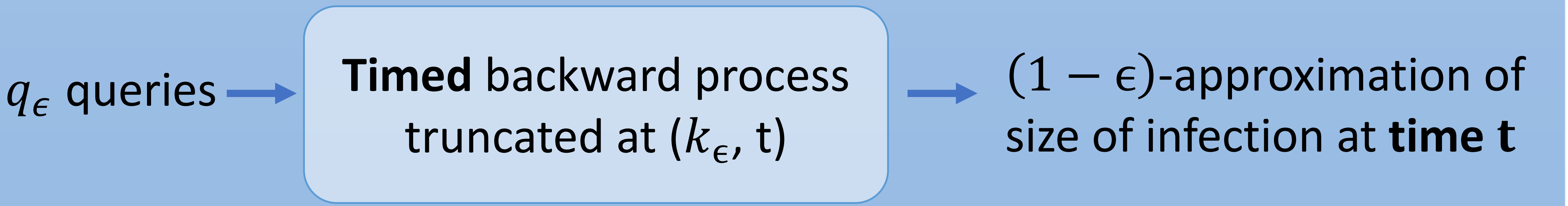
- Their running times are independent of network size!
- They do not assume anything about the structure of the network
- Can be implemented in a way that preserves edge-differential privacy [book by Dwork, Roth ‘14]
- Can be adapted to evolving networks [survey by Hanauer, Henzinger, Schulz ‘21]

Most important:

A constant No. of queries to the local estimators gives a $(1-\epsilon)$-approximation for the final size and time evolution of the epidemics.

Theorem [Alimohmmadi, Borgs, Hofstad, Saberi ('22)]

Consider the SIR process on a *convergent* sequence of graphs, in which each node is infected with probability $\alpha > 0$ and susceptible otherwise. Then for any $\epsilon > 0$, there exist constants $q_\epsilon, k_\epsilon \geq 0$ such that after q_ϵ queries, asymptotically:



q_ϵ queries

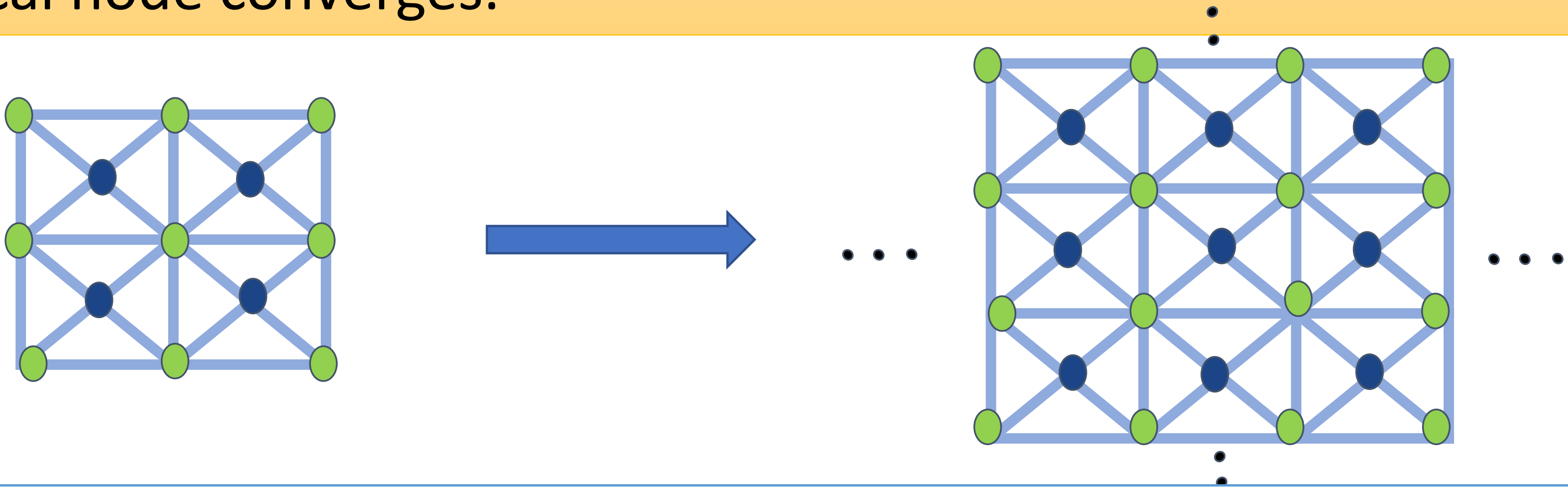
Timed backward process truncated at (k_ϵ, t)

$(1 - \epsilon)$-approximation of size of infection at **time t**

And q_ϵ queries to backward process truncated at (k_ϵ, ∞), gives $(1 - \epsilon)$-approximation of the **final size,** whp.

Graph convergence [Benjamini-Schramm 01]

A sequence of graphs is convergent if the distribution of the neighborhood of a typical node converges.



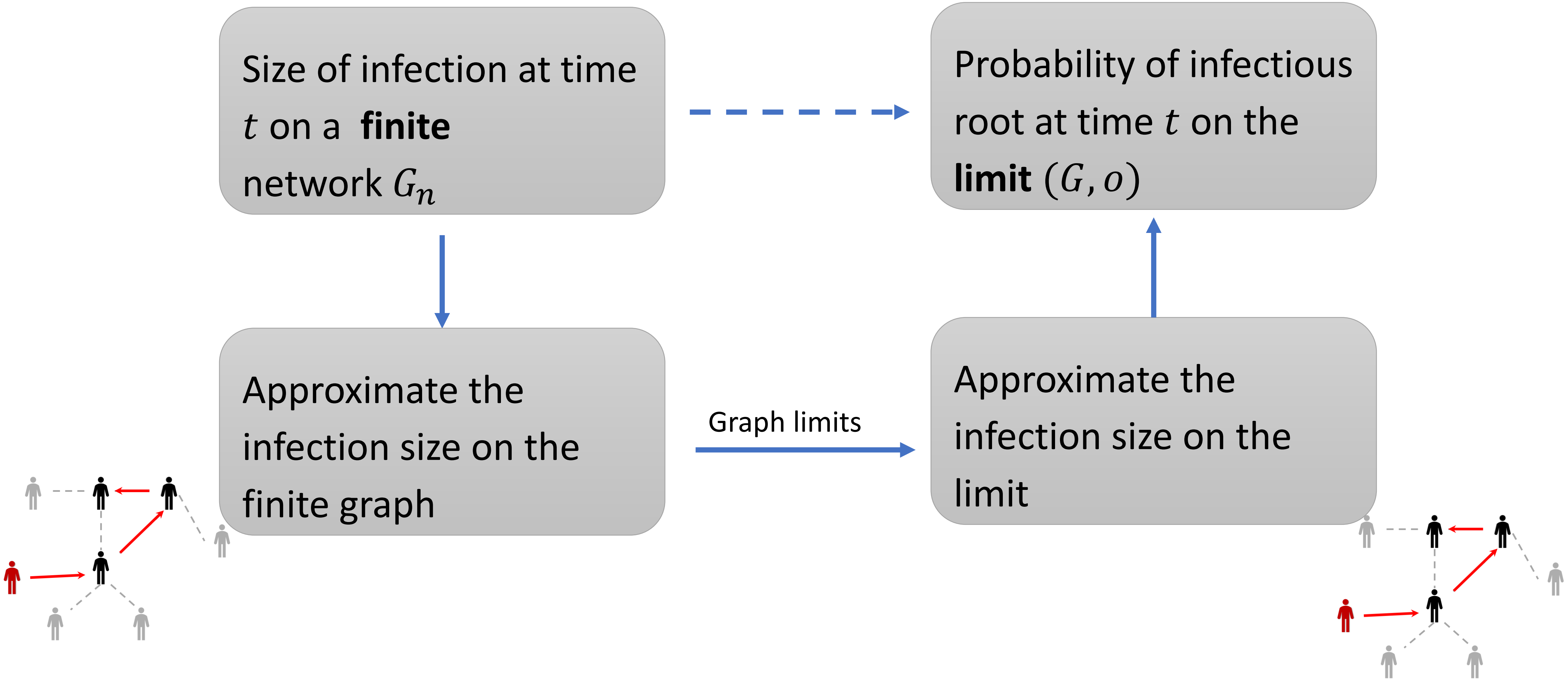
A sequence of finite graphs $\{G_n\}_{n\in}$ converges in probability to $(\mathrm{G},\mathrm{o}) \sim \mu$ if for any rooted graph H and integer k, the probability that a k neighborhood of a random node is isomorphic to H converges,

$$\frac{1}{|V(G_n)|}\sum_{v\in V(G_n)} \mathbb{I}(B_{k(G_n,v)} \simeq H) \xrightarrow{\mathbb{P}} \mathbb{P}_{(G,o)\sim\mu}[\mathrm{B_k}(\mathrm{G},\mathrm{o}) \simeq H]$$

Well-known network models locally converge

- Erdos-Renyi
- Configuration model [Molloy-Reed, Newman-Barabasi-Watts ‘11
- Preferential attachment [Bollobás-Riordan ‘03]
- Small-world networks [Watts-Strogatz ‘00]
- Random geometric graphs [Estrada, Meloni, Sheerin, Moreno ’15]
- Household models [Ball-Sirl-Trapman 2009, Hofstad-Leeuwaarden-Stegehuis. ’15]

Proof Idea: Application of graph limits to epidemics



Size of infection at time t on a **finite** network G_n

Probability of infectious root at time t on the **limit** (G,o)

Approximate the infection size on the finite graph

Graph limits

Approximate the infection size on the limit

Summary so far

- Simple algorithm that uses local information about a small sample of nodes for estimating the final size and the time evolution of the epidemics.
- Correctness does rely on specific features of the graph (e.g. degree sequence, local tree structure, or independence of edges)

On the other hand, the initial condition in our theorem (a constant fraction of nodes infected at the start) is restrictive.

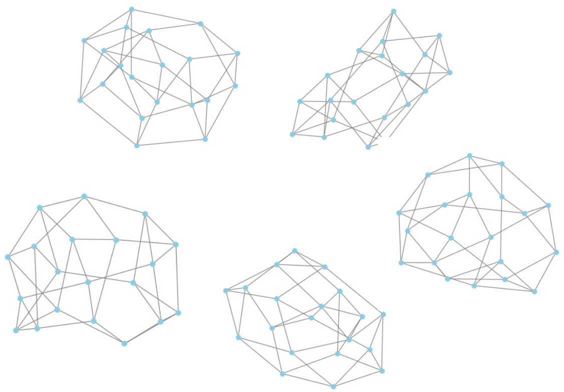
Part II: epidemics starting from a single infection

Epidemics starting from one node

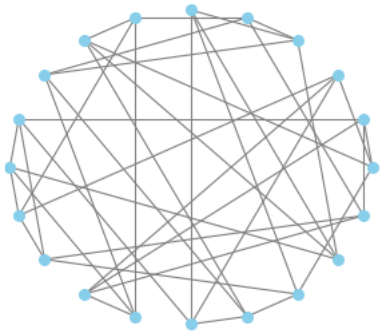
Initially a uniform random node is in I, Everyone else is susceptible.

The problem becomes much more complicated:

- Hard to control the stochasticity
- The infection may die out fast (the probability of outbreak?)
- Same local structure but different outbreak



A collection of 4-regular random graphs, each of size $\log n$



A 4-regular random graph of size n

Epidemics starting from one node

Initially a uniform random node is in I, Everyone else is S.

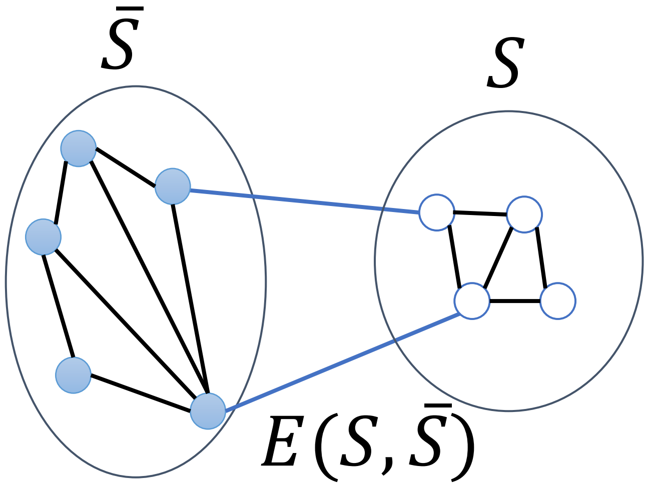
The problem becomes much more complicated:

- We need a condition that the network is well-connected .
- We can only analyze the process when the recovery times are constant.
- We cannot say anything about the time evolution (yet).

Well-connected graphs (expanders)

G is a ϕ-expander if for every subset S $\subset V(G)$

$$E(S,\bar{S}) \geq \phi \min(|S|,|\bar{S}|)$$

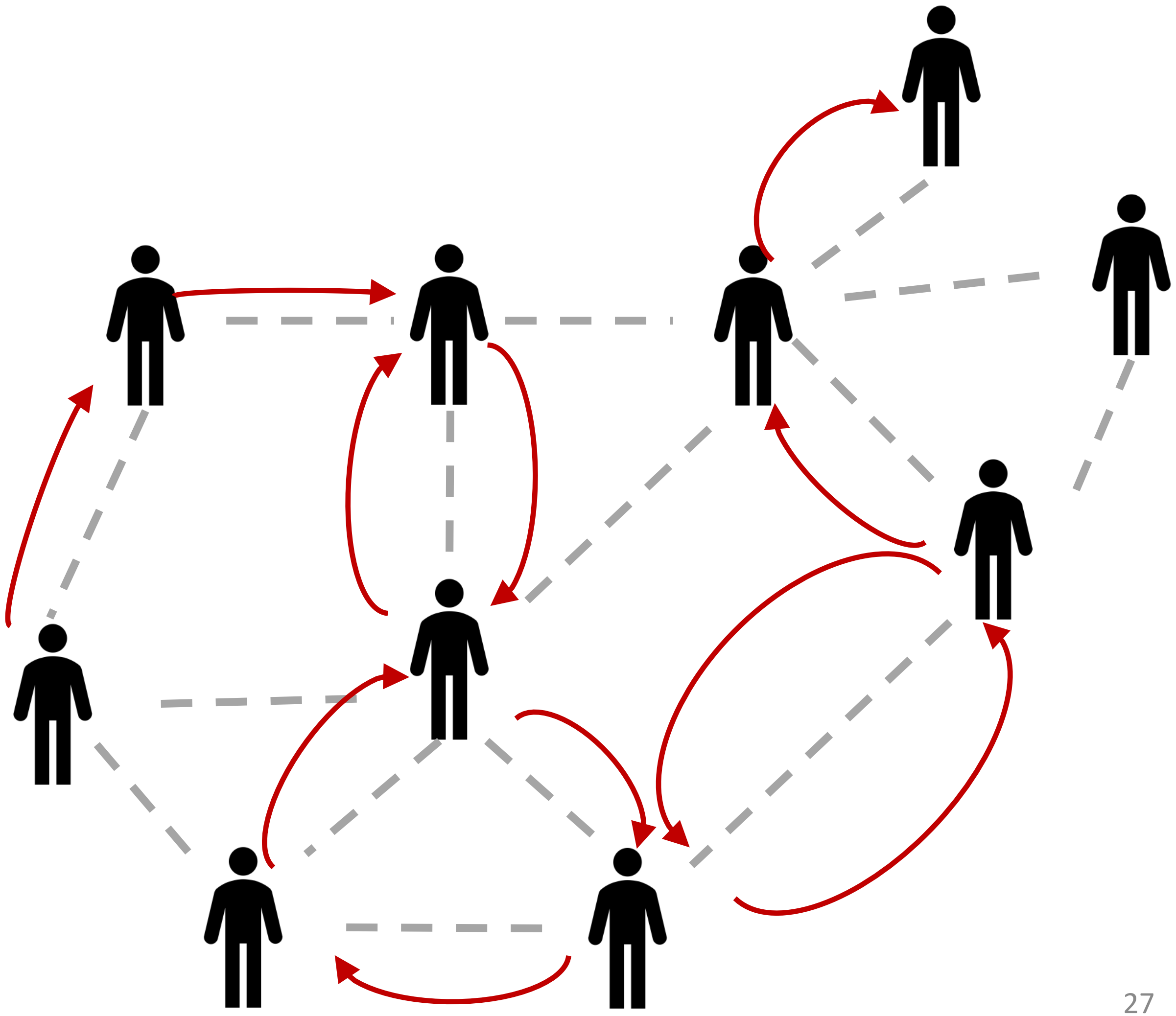


i.e. you can't disconnect a large set of vertices by removing a few edges

SIR with constant recovery time

Transmission network:

- Replace each edge $\{i, j\}$ by directed edges $i \to j$ and j→ i
- Keep directed edges **independently** with probability p



Result at a high level

Consider SIR with constant recovery time with **single seeding** initial condition in well-connected network. Then local information is enough to estimate the relative **size** and **probability** of an outbreak.

Event of outbreak: when $\omega(n)$ people eventually get infected.

Theorem. [Alimohammadi, Borgs, Saberi (AOP’23, SODA’22)]
Let $\{\mathrm{G_n}\}_{\mathrm{n}\in\mathbb{N}}$ be a sequence of well-connected graphs with bounded average degree. Let $R(\infty)$ be final infection size for SIR with constant recovery time. Then there exists p_c and $\zeta(p)$ s.t. for transmission probability $p < p_c$,

$$\frac{R(\infty)}{n} \xrightarrow{\mathbb{P}} 0$$

Infection will die almost surely

and for $p > p_c$:

$$\frac{R(\infty)}{n} \xrightarrow{\mathbb{P}} \chi_{\mathrm{p}}$$

infection dies or infects $\zeta(\mathrm{p})n + o(n)$

$$\chi_{\mathrm{p}} = \begin{cases} 0. & \text{with prob } 1 - \zeta(p) \\ \zeta(p). & \quad\quad \text{with prob } \zeta(p) \end{cases}$$

Theorem. [Alimohammadi, Borgs, Saberi (AOP'23, SODA'22)]
Let $\{\mathrm{G_n}\}_{\mathrm{n}\in\mathbb{N}}$ be a sequence of well-connected graphs with bounded average degree. Let $R(\infty)$ be final infection size for SIR with constant recovery time. Then there exists p_c and $\zeta(p)$ s.t. for transmission probability $p < \ p_c$,

$$\frac{R(\infty)}{n} \xrightarrow{\mathbb{P}} 0$$

Infection will die almost surely

and for $p > \ p_c$:

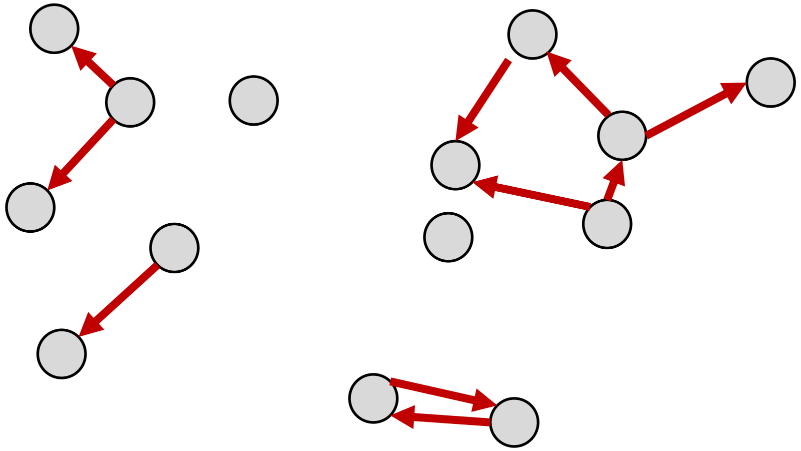
$$\frac{R(\infty)}{n} \xrightarrow{\mathbb{P}} \chi_{\mathrm{p}}$$

infection dies or infects $\zeta(\mathrm{p})n + o(n)$

Furthermore, p_c and $\zeta(p)$ can be estimated using local queries.

Proof: a structure theorem

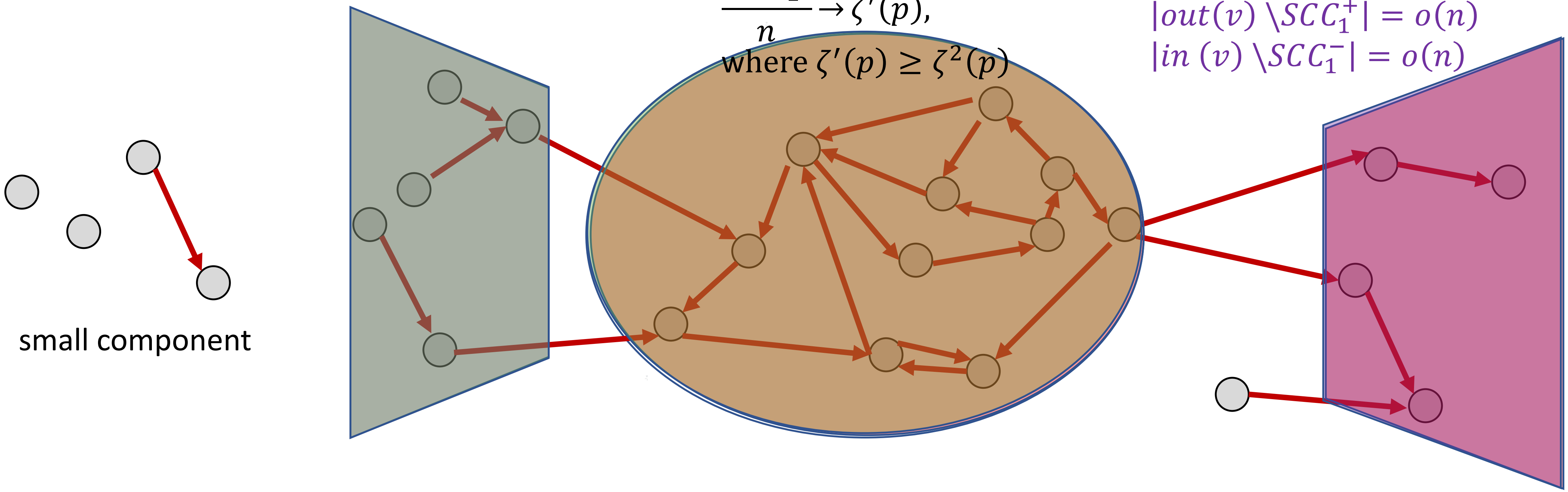
When $p < p_c(G)$ the transmission network is composed of small components



i.e. $\frac{|out(v)|}{n} \xrightarrow{\mathbb{P}} 0, \frac{|in(v)|}{n} \xrightarrow{\mathbb{P}} 0,$ and $\frac{|SCC_1|}{n} \xrightarrow{\mathbb{P}} 0.$

Proof: a structure theorem

When $p > \ p_c(G)$ the transmission network has a ``bow-tie'' structure:



$$\frac{|SCC_1|}{n} \overset{\mathbb{P}}{\rightarrow} \zeta'(p),$$

where $\zeta'(p) \geq \zeta^2(p)$

$|out(v)\ \backslash SCC_1^+| = o(n)$

$|in\ (v)\ \backslash SCC_1^-| = o(n)$

$\boldsymbol{SCC_1^-}$: Nodes leading to an outbreak

$$\frac{1}{n}|SCC_1^+| \overset{\mathbb{P}}{\rightarrow} \zeta(p)$$

$\boldsymbol{SCC_1^+}$: Nodes infected in an outbreak

$$\frac{1}{n}|SCC_1^-| \overset{\mathbb{P}}{\rightarrow} \zeta(p)$$

Theorem [Alimohammadi, Borgs, Saberi (AoP '23)]: for the same models and parameters as in Theorem 2,

- If $p < p_c(G)$, for a uniform random node v whp $\frac{|out(v)|}{n} \xrightarrow{\mathbb{P}} 0, \frac{|in(v)|}{n} \xrightarrow{\mathbb{P}} 0,$ and $\frac{|SCC_1|}{n} \xrightarrow{\mathbb{P}} 0.$

- If $p > p_c(G)$:
  - ○ There exists $\zeta'(p) \geq \zeta^2(p)$ such that $\frac{|SCC_1|}{n} \xrightarrow{\mathbb{P}} \zeta'(p)$.
  - ○ $\frac{1}{n}|SCC_1^+| \xrightarrow{\mathbb{P}} \zeta(p)$ and $\frac{1}{n}|SCC_1^-| \xrightarrow{\mathbb{P}} \zeta(p)$
  - ○ For a uniform random node v whp $|out(v) \setminus SCC_1^+| = o(n)$, and $|in\ (v) \setminus SCC_1^-| = o(n)$.

A few notes

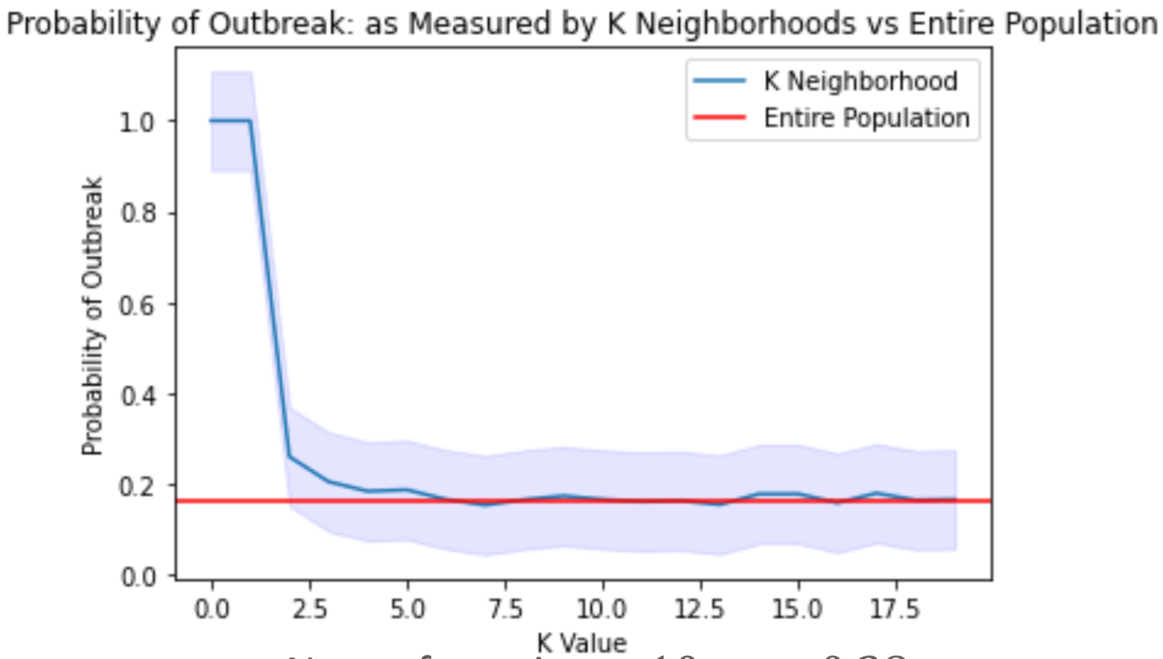
- Bow-tie structure coined by Broder et al:
  Web is a bow-tie [Broder, Kumar, Maghoul, Raghavan, Rajagopalan, Stata, Tomkins, Wiener (2000)]

- Similar structure theorem was proved Bow-tie on Erdos Renyi [Karp '90 Luzcak '90] and configuration model [Cooper, Frieze '04]. Similar results were known for preferential attachment graphs and small-world networks.

- Our result unifies and (weakly) generalizes them.

Does the Algorithm Work on Real-world Graphs?

- Copenhagen dataset
- Bluetooth data of 700 students
- Edge exists if distance <6 ft



Num of queries $= 10, p = 0.28$,
95% confidence interval is highlighted

Data: “Interaction data: Copenhagen Networks Study”
[Sapiezynski, Stopczynksi, Lassen, & Lehman, Nature ’19]

Takeaways: Local Information Goes a Long Way!

Initial condition: well-mixed seeding
Local information is enough to estimate the **time evolution** of the epidemics.

Initial condition: single seeding
Local information is enough to estimate the **probability** and relative **size** of an **outbreak** for large class of networks under a simple infection spread**.**

Other applications of graph limits in analyzing global quantities with local structures: Graph Neural Networks, network games

References

Alimohammadi, Borgs, Saberi, "*Algorithms Using Local Graph Features to Predict Epidemics*" (SODA' 2022)

Alimohammadi, Borgs, Saberi, "*Locality of Random Digraphs on Expanders*" (Annals of Probability, 2022)

Alimohammadi, Borgs, van der Hofstad, Saberi, "*Epidemics on Networks is Local.*" (working paper)

Thank You

yeganeh@stanford.edu

Relative Size of the Giant in Expanders

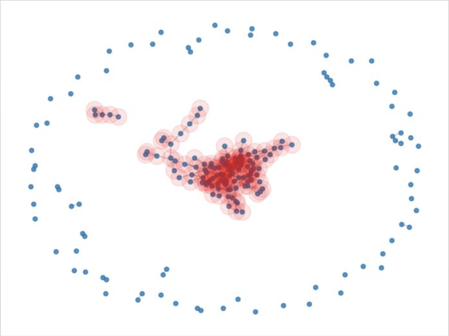
Theorem 1. [Alimohammadi, Borgs, Saberi ’21 (Annals of Probability)]
Let $\{\mathrm{G_n}\}_{\mathrm{n}\in\mathbb{N}}$ be a sequence of convergent large-set expanders with bounded. Let $\mathrm{C_i}$ be the i^{th} largest component. If $\mathrm{p} \neq \ \mathrm{p_c}(\mu)$,

$$\frac{|\mathrm{C_1}|}{\mathrm{n}} \overset{\mathbb{P}}{\rightarrow} \zeta(\mathrm{p}),$$

Also for all $p \in [0,1], \frac{|\mathrm{C_2}|}{\mathrm{n}} \overset{\mathbb{P}}{\rightarrow} 0.$

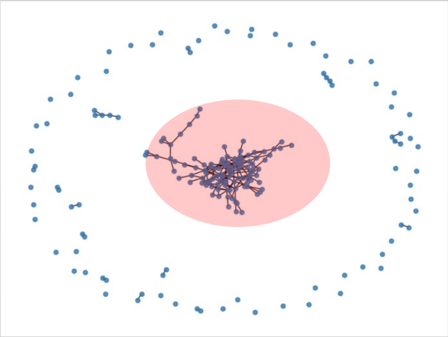
Takeaway: Giant in convergent expanders is unique, and its size converges to its limit.

Proof Sketch: Size of the Giant Converges



Relative num of nodes in a component larger than k in $G(p)$ $\to^{P}$ $\zeta_k(p)$

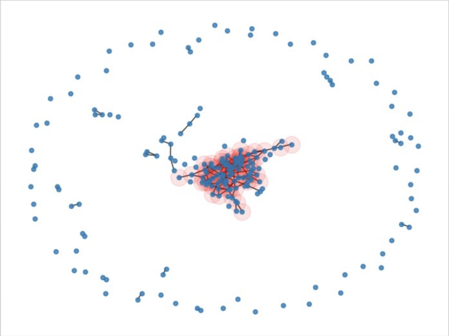
≥



Relative num of nodes in the largest connected component of $G(p)$

≥

Sprinkling!

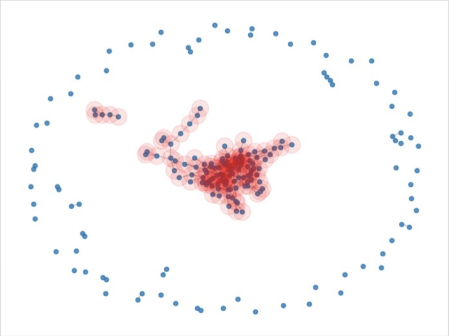


Relative num of nodes in a component larger than k in $G(p-\epsilon)$ $\to^{P}$ $\zeta_k(p-\epsilon)$

$$\zeta_k(\mathrm{p}) := \mathbb{E}_{(G,o)\sim\mu}[\mathbb{P}_{G(p)}(|\text{connected component of } o| \geq k\ )].$$

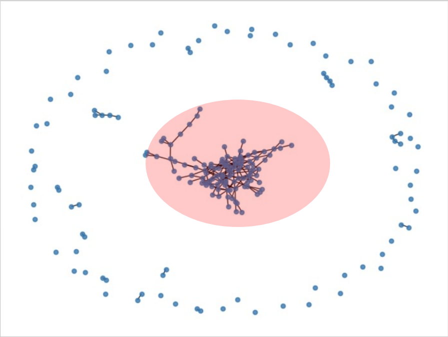
$$\lim_{k\to\infty} \zeta_k(\mathrm{p}) = \zeta(p).$$

Proof Sketch: Size of the Giant Converges



Relative num of nodes in a component larger than k in $G(p)$ $\to^P \zeta_k(p)$

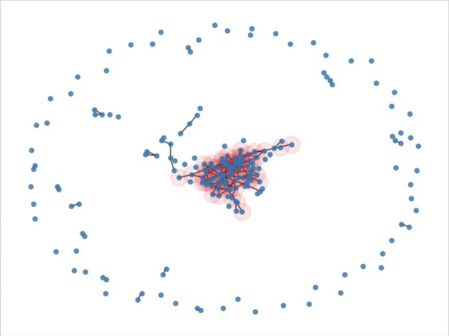
≥



Relative num of nodes in the largest connected component of $G(p)$

≥

Sprinkling!



Relative num of nodes in a component larger than k in $G(p-\epsilon)$ $\to^P \zeta_k(p-\epsilon)$

Lemma. For a sequence of graphs satisfying the assumptions of Theorem 2, $\zeta(p)$ is continuous for all $p \neq p_c(\mu)$. Equivalently, the limit μ is ergodic.

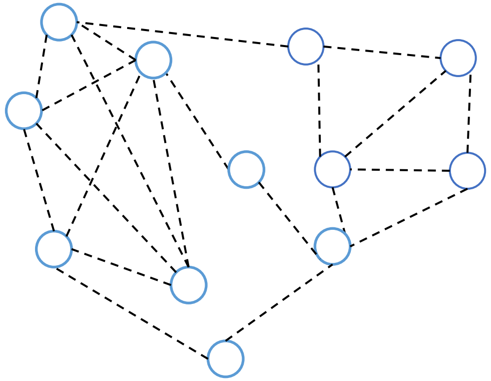
Brief History of Sprinkling

[Erdös, Rényi’60]
[Posa’76][Ajtai, Kolmós, Szemerédi ‘82]
[Bollobás, Riordan ‘01] [Alon, Benjamini, Stacey ‘02]
[Borgs, Chayes, van der Hofstad, Slade, Spencer ‘07]
[Benjamini, Nachmias, Peres ‘09]
[Janson, Rucinski’10] [van der Hofstad, Nachmias ‘17]
[Krivelevich, Sudakov ’17]
[Dudek, C. Reiher, A. Rucínski, and M. Schacht ‘20]
[Nenadov, Trujic ’21][Easo, Hutchcroft ’21]

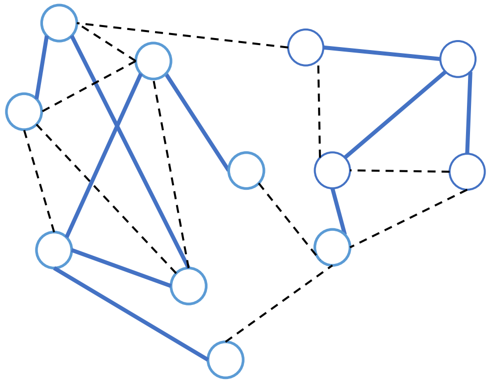
[Alimohammadi, Borgs, Saberi ‘21+]

Proof: the Lower Bound

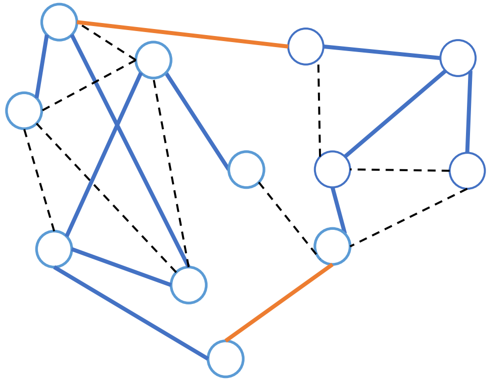
Step 0: For some $\epsilon > 0$ let $p_1 = p_c(\mu) + \epsilon$ be such that $1 - p = (1 - p_1)(1 - \epsilon)$.
Consider two copies of percolation $G_n(p_1)$ and $G_n(\epsilon)$. The union of them gives an instance of $G_n(p)$.



The original graph G_n



$G_n(p_1)$



$G_n(\epsilon)$

Proof: the Lower Bound

Step 0: For some $\epsilon > 0$ let $p_1 = p_c(\mu) + \epsilon$ be such that $1 - p = (1 - p_1)(1 - \epsilon)$.
Consider two copies of percolation $G_n(p_1)$ and $G_n(\epsilon)$. The union of them gives an instance of $G_n(p)$.

Step 1: There exists some $\delta > 0$ such that for all $K > 0$, whp there are δn nodes with component larger than K in $G_n(p_1)$.

Step 2 (Sprinkling): Let $\mathrm{Z_K} = \{\text{nodes with component larger than } K\}$.
There is a path in $G_n(\epsilon)$ between any two large partition of components in Z_K:

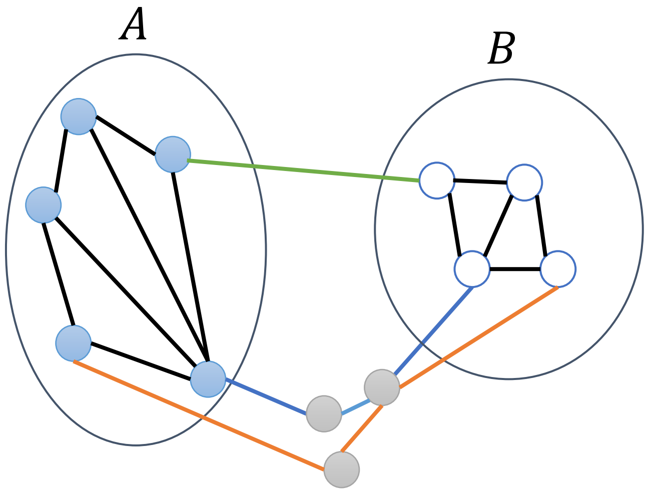
$$\mathbb{P}_{G_n(\epsilon)}\left(\exists\, A, B \subseteq 2^{Z_K}\text{: } A\text{ , }B \text{ disconnected in } G_n(\epsilon) \text{ and } G_n(p_1),\ |\mathrm{A}|,|B| \geq \frac{\delta n}{3} \mid G_n(p_1)\right)$$
$$\leq \exp(-nc_{\{\alpha,\delta,d,\epsilon\}})$$

Step 3: $\mathbb{P}_{G_n(p)}\left(\mathit{contains\ a\ component\ of\ size\ } \frac{\delta n}{3}\right) \to 1,$ as $n \to\ \infty$.

Step 2: Sprinkling

Step 2 (Sprinkling): There is a path in $G_n(\epsilon)$ between any two large partition of components in Z_K:

$$\mathbb{P}_{G_n(\epsilon)}\left(\exists\, A,B \subseteq 2^{Z_K}\text{: } A\,,B \text{ disconnected in } G_n(\epsilon) \text{ and } G_n(p_1),\ |\mathrm{A}|,|B| \geq \frac{\delta n}{3} \mid G_n(p_1)\right) \leq \exp(-nc_{\{\alpha,\delta,d,\epsilon\}})$$



Menger's Theorem. Let G be a finite undirected graph and A and B two disjoint set of vertices. Then the minimum edge-cut between A and B is equal to the number of pairwise edge-independent paths from A to B.

There are $\frac{\delta\alpha n}{3}$ edge-disjoint paths in G_n between A and B (expansion).

Since the average degree is bounded by d, the length of half of these paths is bounded by $\ell = \frac{6d}{\delta\alpha}$. (# paths = $\frac{\delta\alpha n}{6}$)

Each path appear in $G_n(\epsilon)$ with probability ϵ^{ℓ}.

The probability that non of the paths appear in $G_n(\epsilon)$: $\left(1-\epsilon^{\ell}\right)^{\#paths}$

Number of $A\,,B$ partitions in $G_n(p_1)$: $2^{\frac{\mathrm{n}}{K}}$

Finally: $2^{\frac{n}{K}}\left(1-\epsilon^{\frac{6d}{\delta\alpha}}\right)^{\frac{\delta\alpha n}{6}} \leq \exp\left(n(\frac{1}{K}-\frac{\delta\alpha}{6}\epsilon^{\frac{6d}{\delta\alpha}})\right)$