

# Impact of graph perturbations on structural and dynamical properties

---

*Anil Kumar Vullikanti*

Dept of Computer Science and Biocomplexity Institute of Virginia Tech

Joint work with: Abhijin Adiga, Chris Kuhlman, Henning Mortveit

March 22, 2016

# Networks and their applications

---

## Typical workflow

- Develop network model of data: infer nodes, edges and associated attributes
- Using structural properties
  - Identify nodes based on their attributes (e.g., centrality, core number)
  - Find subset of nodes with some characteristics (e.g., community,  $k$ -core)
- Dynamics on networks
  - Model dynamical process on network: need to infer node states, functions determining state changes
  - Identify nodes which have a significant impact on dynamics

What happens if the networks are not accurate or incomplete, e.g., due to errors in inference?

# Sources of uncertainty in networks

---

- Internet router graph
  - Graph on the set of routers
  - Inferred by traceroutes and other indirect techniques
- Biological networks
  - Gene/protein networks
  - Inferred based on experimental correlations
- Social networks
  - Twitter mentions graph (who mentions whom)
  - Sampling using twitter APIs

Inherently noisy: missing/wrong nodes/edges, sampling biases



# How to model uncertainty?

- No right model: relevance depends on application and methodologies used to construct the network.
- Simplest model: stochastic perturbation
  - Pick random pair of nodes  $u, v$  with probability  $P_G(u, v)$  and add/delete  $(u, v)$  with probability  $\epsilon$
  - $G(\epsilon)$ : perturbed graph,  $\epsilon =$  the perturbation factor.
  - **Uniform Perturbation** (ERP model):  $P_G((u, v)) = 1/n \Rightarrow$  expected #edges altered  $\approx \frac{\epsilon n}{2}$
  - **Degree Assortative Perturbation** (CLP):  $P_G((u, v)) \propto d(u) \cdot d(v)$ , where  $d(u)$  is the degree of node  $u \Rightarrow$  Expected #edges altered  $\approx \epsilon m$ , where  $m =$  #edges
  - **Link Prediction Based Model** (LPP): Pick edges predicted by link prediction models, e.g., [Clauset et al., 2008]
- **Sampled subgraphs**: random subgraph  $G_p$  obtained by picking each edge in  $p$  independently with probability  $p$ 
  - Networks from online social media are usually sampled

# Related work (partial)

---

- Sensitivity analysis for social network measures
  - Sensitivity of centrality measures under stochastic perturbations [Borgatti et al., 2006]
  - Effect of random rewirings, e.g., [Watts, Strogatz]
- Rigorous analysis of the impact of stochastic perturbations on structural properties
  - Graph diameter [Flaxman and Frieze, 2004]: for any  $\epsilon > 0$ , the diameter of the perturbed graph  $G(\epsilon)$  becomes  $O(\log n)$ , with high probability.
  - Graph expansion [Flaxman, 2007]: if “small” sets have expansion, then  $G(\epsilon)$  has high expansion, with high probability.
  - Smoothed analysis
- Effect of perturbations on dynamical properties: [He et al., KDD 2014], [Lahiri et al., 2008]

# Focus of this talk

---

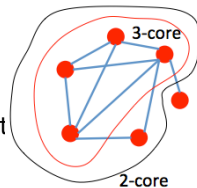
Study effect of network uncertainty on

- Core decomposition in graphs
- Dynamical properties in Linear Threshold (LT) and Independent Cascades (IC) models



# Core decomposition: definitions

- $C_k(G)$ :  $k$ -core of graph  $G = (V, E)$  is the maximal subgraph in which each node has degree at least  $k$ 
  - Can be determined by repeatedly removing nodes of degree less than  $k$
- Core-number of a node  $v$ : largest  $k$  such that  $v$  is in the  $k$ -core of  $G$
- $S_k(G)$ :  $k$ -shell = subset of nodes with core number equal to  $k$



# Core decomposition: applications and questions

---

Popular measure in network analysis:

- High core set is generally well-connected, useful for controlling dynamical properties (e.g., securing selected nodes, or selecting as sources for influence maximization)
- Typical uses: compute core decomposition and identify nodes with high core number

## Main questions

- Are the top core sets and the core decomposition robust under perturbations or sampling?
- Can we determine regimes where the top core sets are “stable”?

# Definitions

---

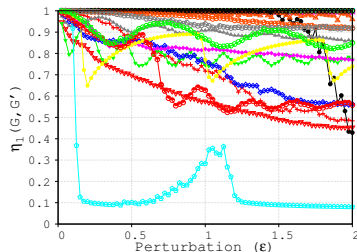
- $k_{max}(G)$ : maximum core number in  $G$
- Consider graph  $G'$  resulting from perturbations ( $G' = G(\epsilon)$ ) or sampling ( $G' = G_p$ )
- $\eta_j(G, G')$ : Jaccard similarity of the top  $j$ -core sets in graphs  $G$  and  $G'$ , i.e., between  $C_{k_{max}(G)-j+1}$  and  $C_{k_{max}(G')-j+1}$
- $\delta(G, G')$ : variation distance between shell size distributions  $\langle |S_1(G)|, \dots, |S_{k_{max}(G)}(G)| \rangle$  in  $G$  and  $\langle |S_1(G')|, \dots, |S_{k_{max}(G')}(G')| \rangle$  in  $G'$ .

Class	Network	N	E	$k_{\max}$	$ C_{k_{\max}}(G) $
Autonomous Systems	As20000102	6474	12572	12	21
	Oregon1010331	10670	22002	17	32
	Oregon2010331	10900	31180	31	78
Co-authorship	Astroph	17903	196972	56	57
	Condmat	21363	91286	25	26
	Grqc	4158	13422	43	44
	Hepph	11204	117619	238	239
	Hepth	8638	24806	31	32
Citation	HepPh	34546	420877	30	40
	HepTh	27770	352285	37	52
Communication	Email-EuAll	265214	364481	37	292
	Email-Enron	33696	180811	43	275
Social	Epinion	75877	405739	67	486
	Slashdot0811	77360	469180	54	129
	Soc-Slashdot0902	82168	504230	55	134
	Twitter	22405	59898	20	177
	Wiki-Vote	7066	100736	53	336
	Twitter "mentions"	2616396	4677321	19	210
Internet peer-to-peer	Gnutella04	10876	39994	7	365
	Gnutella24	26518	65369	5	7480
Synthetic graphs	Regular ( $d = 20$ )	10000	100000	20	10000

Different real-world and synthetic graphs used in our experiments and their properties.

# Results (I): sensitivity of the core-decomposition to noise

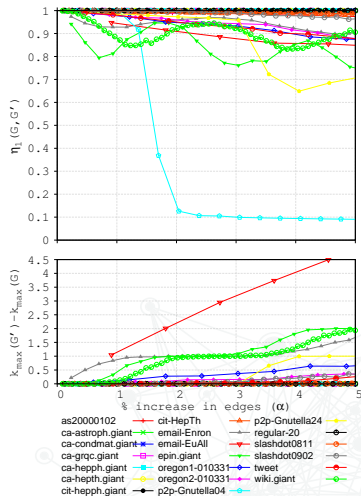
- Shell size distribution sensitive to  $\epsilon$  in the case of unbiased perturbations
  - $G(\epsilon)$  always has a 2-core with high probability
- Effect on Jaccard index  $\eta_k(G, G(\epsilon))$  of top cores in  $G(\epsilon)$  and  $G$ , as a function of  $\epsilon$ 
  - $\eta_k$  stable to unbiased perturbations, but sensitive to degree-biased perturbations
  - Non-monotone variation in 20 diverse real and random networks
  - Citation networks are much more stable than influence and P2P networks



as20000102	cit-HepTh	p2p-Gnutella24	
ca-astroph.giant	email-Enron	regular-20	
ca-condmat.giant	email-EuAll	slashdot0811	
ca-grqc.giant	epin.giant	slashdot0902	
ca-hep-ph.giant	oregon1-010331	tweet	
ca-hep-th.giant	oregon2-010331	wiki.giant	
cit-hep-ph.giant	p2p-Gnutella04		

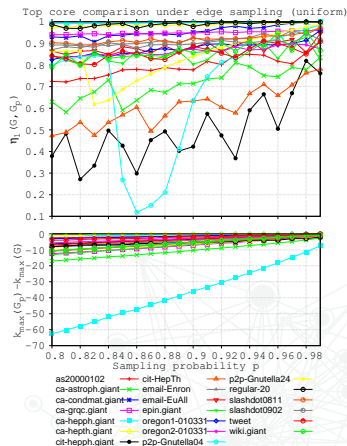
# Results (II): sensitivity of the core-decomposition to noise

- Jumps in  $\eta_k$  correspond to changes in largest core number
  - Might be possible to identify regions of sensitivity if we can determine whether the largest core number would change
  - Motivates CorePerturbation problem ( $CP(G, E_A, p, k)$ ): determine probability that if there is a  $k$ -core, after  $G$  is perturbed by adding random  $p$  edges from  $E_A$ , each with probability  $p$
  - Turns out to be #P-complete
- Can effects be mitigated by considering top  $k$  cores, with  $k > 1$ ?
  - Sensitivity diminishes as  $k$  increases, however size of  $k$ -core increases
  - $\eta_k$  varies non-monotonically with  $k$



# Results (III): sensitivity of the core-decomposition to sampling

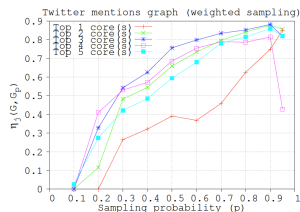
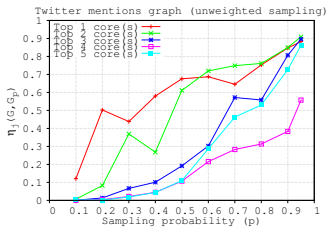
- High sensitivity of  $\eta_k(G, G_p)$  to sampling probability  $p$ 
  - $\eta_k$  is a *non-monotone* function of  $p$  and  $k$  in diverse real and random networks
  - Identifying 80% of the nodes in the top core set requires  $p > 0.6$  in many networks
- Citation networks very sensitive to sampling (quite robust to noise)
- $k_{\max}(G_p)$  scales with  $p$ : can prove that for any constant  $\delta \in (0, 1)$ ,  $k_{\max}(G_p) > (1 - \delta)k_{\max}(G)p/2$ , with high probability, if  $k_{\max}(G) \rightarrow \infty$  as  $n \rightarrow \infty$ .



# Results (IV): sensitivity of the core-decomposition to sampling

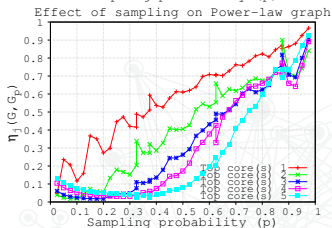
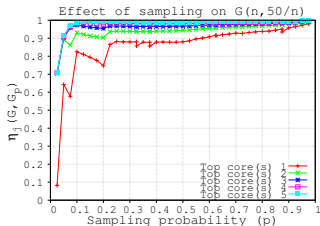
## Twitter mentions graph

- $\eta_k$  is quite low and non-monotone unless sampling probability high
- Higher  $k$  reduces non-monotonicity
- Consider weight of edge = number of mentions
- Consider sampling proportional to edge weights: shows more robust behavior
- Need fairly high sampling to recover large fraction of top cores



# Results (V): sensitivity of the core-decomposition to sampling

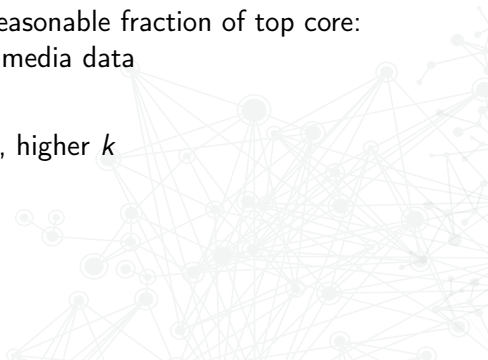
- Inherent aspect of sampling: also occurs in graphs from the Erdős-Rényi and Chung-Lu random graph models
- There exist constants  $c$  and pairs  $p_1, p_2$ , where  $0 < p_1 < p_2 < 1$  such that for  $G \in \mathcal{G}(n, c/n)$ ,  $\eta_1(G, G_{p_1}) > \eta_1(G, G_{p_2})$ , with high probability. (using result of [Pittel et al., 1996] on thresholds for  $k$ -cores in random graphs)



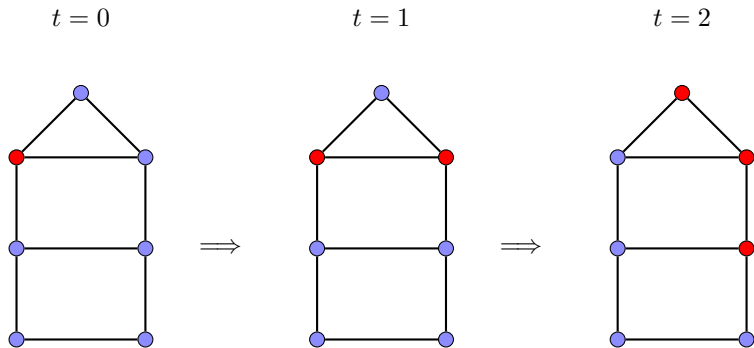
# Implications

---

- Top cores show significant sensitivity to perturbations and sampling
- Non-monotone effects  $\Rightarrow$  careful sensitivity analysis is necessary when using the core structure
- Hardness of CorePerturbation suggests quantifying the effects of uncertainty can be challenging
- High sampling rate needed to recover reasonable fraction of top core: important for networks based on social media data
- Inherent aspect of sampling
- More robust notions: weighted versions, higher  $k$



# Specific models for diffusion

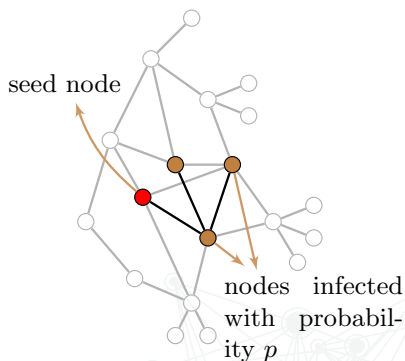


- Nodes in state 0 or 1
- Switch state from 0 to 1, depending on neighbors
- Initially: set of seed nodes infected
- Two specific models: Independent Cascades (IC) and Linear Threshold (LT)

# Independent cascades (IC) model

An infected node  $v$  ( $x_v = 1$ ) infects its neighbors with transmission probability  $p$ . Then, it is removed from the system.

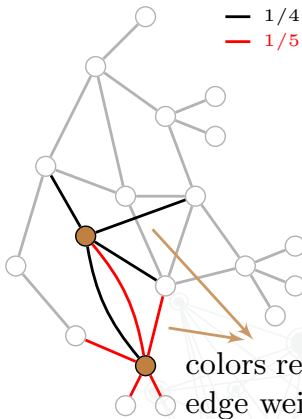
- monotone, stochastic system
- special case of the SIR model
- Throughout we assume that  $G$  is undirected.



# Linear threshold model

- Each node  $v$  has threshold  $\Theta_v \in [0, 1]$ , chosen uniformly at random.
- $v$  is influenced by neighbor  $w$  according to weight  $b_{vw}$  such that  $\sum_{w \in N(v)} b_{vw} \leq 1$ .
- $v$  becomes infected if  $\sum_{w \in N(v)} b_{vw} x_w \geq \Theta_v$ .

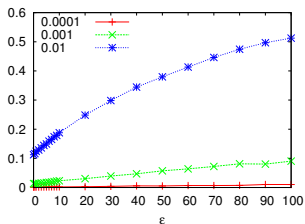
We consider the special case where  $b_{vw} = \frac{1}{d(v)}$ ,  $\forall w \in N(v)$ .



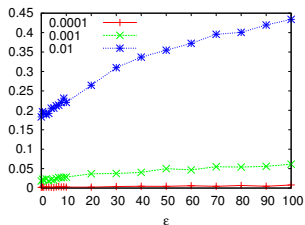
# Our results: $E[\#infections]$ vs $\epsilon$

## Linear Threshold model

Different #seeds



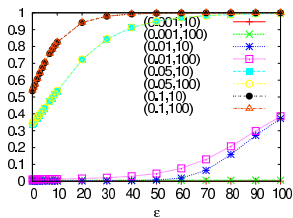
Slashdot network



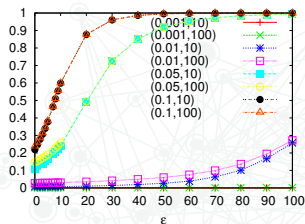
Wiki network

## Independent Cascades model

Different  $p$  and #seeds



Astrophysics citation network

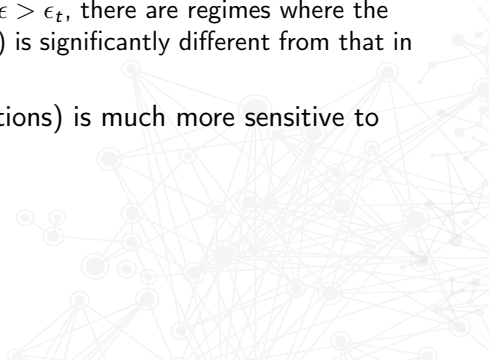


Epinions network

# Our results (summary)

---

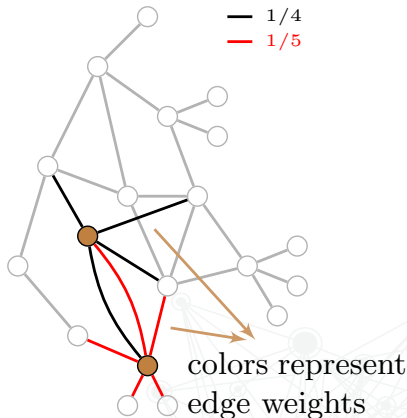
- Variation of  $E[\#\text{infections in } G(\epsilon) | \text{a random initial infection}]$  with  $\epsilon$  in the IC and LT models.
  - In general, behavior very messy, depends on regimes
  - Linear threshold model: Expected number of infections varies “smoothly” with  $\epsilon$ , especially when  $\#\text{seeds}$  small.
  - IC model: The expected number of infections varies “smoothly” with  $\epsilon$ , unless it is in a critical range  $\epsilon_t$ . For  $\epsilon > \epsilon_t$ , there are regimes where the expected number of infections in  $G(\epsilon)$  is significantly different from that in  $G$ .
- Transient behavior (time series of infections) is much more sensitive to perturbations in both models



# Linear threshold model

- Each node  $v$  has threshold  $\Theta_v \in [0, 1]$ , chosen uniformly at random.
- $v$  is influenced by neighbor  $w$  according to weight  $b_{vw}$  such that  $\sum_{w \in N(v)} b_{vw} \leq 1$ .
- $v$  becomes infected if  $\sum_{w \in N(v)} b_{vw} x_w \geq \Theta_v$ .

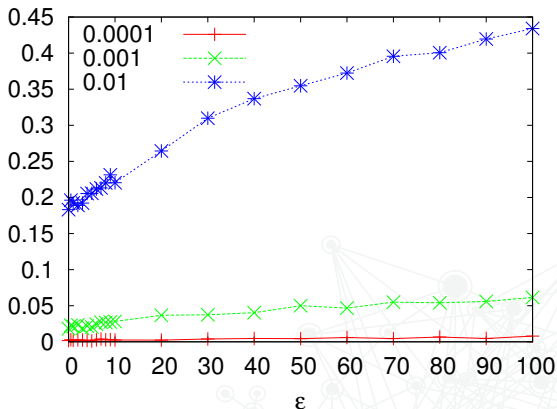
We consider the special case where  $b_{vw} = \frac{1}{d(v)}$ ,  $\forall w \in N(v)$ .



# LT model: effect of random edge additions

**Main result:** Let  $G(\epsilon)$  be the perturbed graph. For a random single seed,  $E[\#\text{infections in } G(\epsilon)] = E[\#\text{infections in } G] + O(\epsilon \log n)$ .

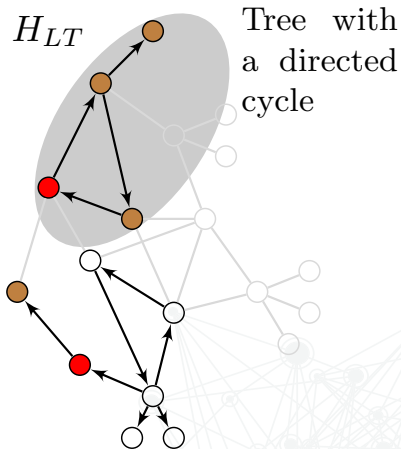
$E[\#\text{infections}]$  is not sensitive to noise in the LT model for small #seeds.



# Proof outline

LT model is equivalent to the following directed percolation process: For each  $v$ , choose a neighbor  $w$  and draw a directed edge  $(w, v)$ .

- Every vertex has exactly one incoming edge.
- Each component is a tree with one directed cycle.
- All vertices reachable from the seed are infected.



# Proof outline

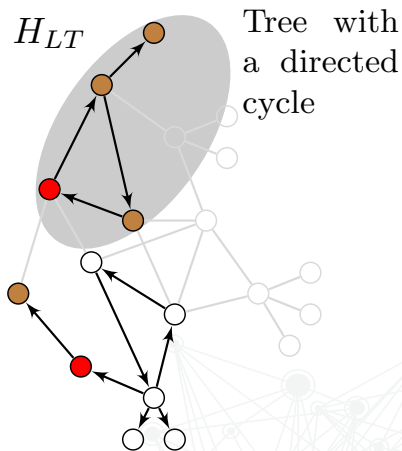
Given a component  $C$  in  $H_{LT}$ :

- Let  $T$  be a directed tree obtained by removing one of the edges of the cycle in  $C$ .
- The depth of  $T$  is  $O(D \log n)$  a.s.

Let  $N(v, T)$  be the number of nodes reachable from  $v$  in  $T$ .

For a random seed in  $C$ , the

$$E[\#\text{infections in } C] = \frac{1}{|T|} \sum_v N(v, T) \leq 2(\text{depth of } T) = O(D \log n)$$



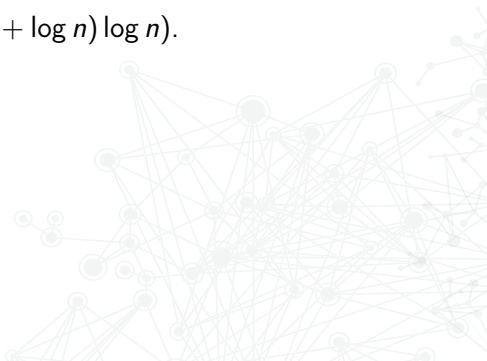
# Proof outline

---

Maximum degree of  $G$  is  $D \Rightarrow E[\#\text{infections in } G] = O(D \log n)$ .

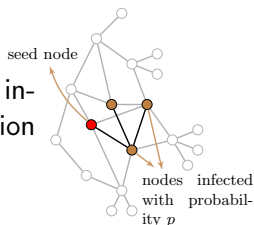
Maximum degree of  $G(\epsilon)$  is  $\leq D + \epsilon + O(\log n)$  a.s.

Hence,  $E[\#\text{infections in } G(\epsilon)] = O((D + \epsilon + \log n) \log n)$ .



# Independent cascades model

An infected node  $v$  ( $x_v = 1$ ) independently infects each of its neighbors with transmission probability  $p$  in one time step.



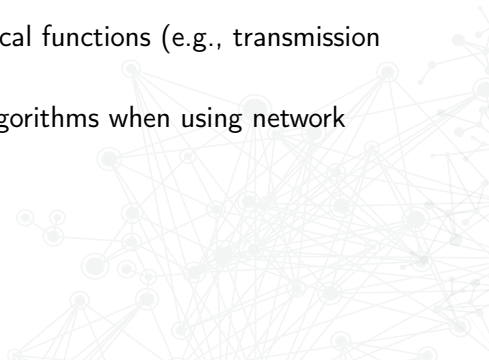
**Main result:** Let  $G(\epsilon)$  be the perturbed graph. Assuming  $G$  exhibits a “phase transition”, there exists a threshold  $\epsilon_t$  (function of  $p$ ) and constants  $\delta < 1$ ,  $c_1 < 1$  and  $c_2 > 1$  such that for  $p < \delta p_c$ :

- 1 if  $\epsilon < c_1 \epsilon_t$ : expected #infections in  $G(\epsilon)$  is not very sensitive to  $\epsilon$ .
- 2 if  $\epsilon > c_2 \epsilon_t$ : expected #infections in  $G(\epsilon)$  is much larger than in  $G$ .

# Conclusions

---

- Uncertainty is an important issue in network abstractions
- Use structural properties carefully
- Impact on dynamics is very model dependent
  - SIR vs linear threshold
  - Ongoing work:  $t$ -threshold models behave differently
  - Transient behavior more sensitive
- Dynamics sensitive to other aspects: local functions (e.g., transmission probability)
- Implications: need to develop robust algorithms when using network abstractions



Thank you!

