CSC 591/791: Graph Data Mining: Theory, Algorithms and Applications

Virus Propagation on Networks

Guest Lecture

Gonzalo Bello

Oak Ridge National Laboratory

NC STATE UNIVERSITY
Department of Computer Science

Outline

- Virus Propagation: Introduction.
- Virus Propagation Models (VPMs).
- Virus Propagation on Static Contact Networks.
  - Epidemic Threshold.
  - Effective Strength.
- Virus Propagation on Alternating Contact Networks.
- Immunization Policies.
  - *NetShield* Algorithm.
- Bonus Problem: Viral Marketing.

Introduction

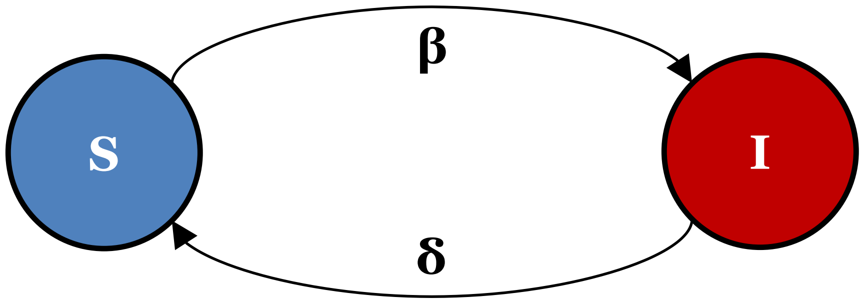
- Fundamental questions in **epidemiology**:
  - Will a virus result in an epidemic or will it die quickly?
  - Given a limited number of vaccines available, which individuals should be immunized to prevent an epidemic?
- Recent works on **virus propagation** have approached these problems from a network perspective.
- Other applications:
  - Minimize the spread of malware in a computer network.
  - Maximize the spread of a marketing campaign across a social network.

Virus Propagation Model (VPM)

- A **VPM** is a simplified model of disease spread that provides general information about the behavior of a disease:
  - How virulent is the disease?
  - How quickly does the host recover (if ever)?
  - Does the host obtain (or is born with) immunity?
  - How quickly does the host lose immunity (if ever)?
- Some common VPMs:
  - SIS, SIR, $SIRS$, $SEIV$.

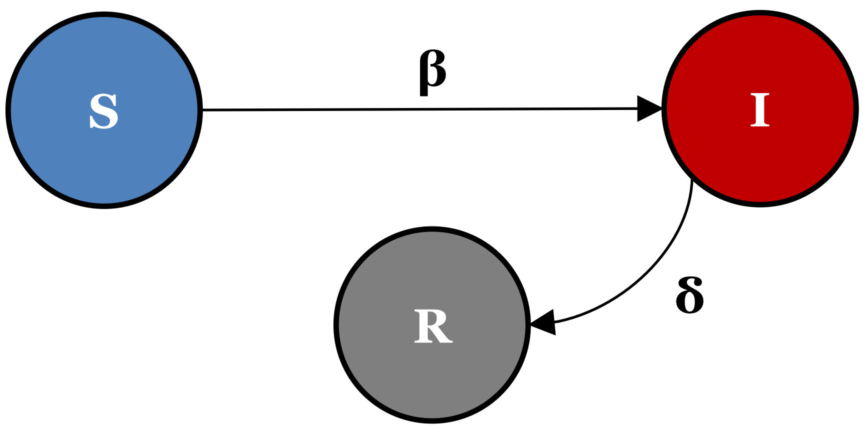
***SIS* VPM**

- ***SIS* VPM** (“susceptible, infected, susceptible”):
  - Example: flu.
  - Two states: **Susceptible** (S), **Infected** (I).
  - Transition probabilities: attack/transmission probability β, healing probability δ.
  - No immunity.



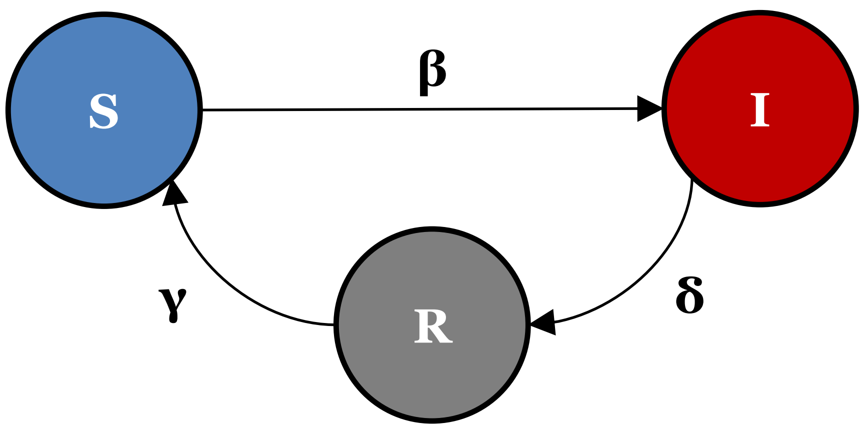
SIR VPM

- ***SIR* VPM** (“susceptible, infected, recovered”):
  - Example: mumps.
  - Three states: **Susceptible** (S), **Infected** (I), **Recovered** (R).
  - Transition probabilities: attack/transmission probability β, healing probability δ.
  - Life-time immunity.



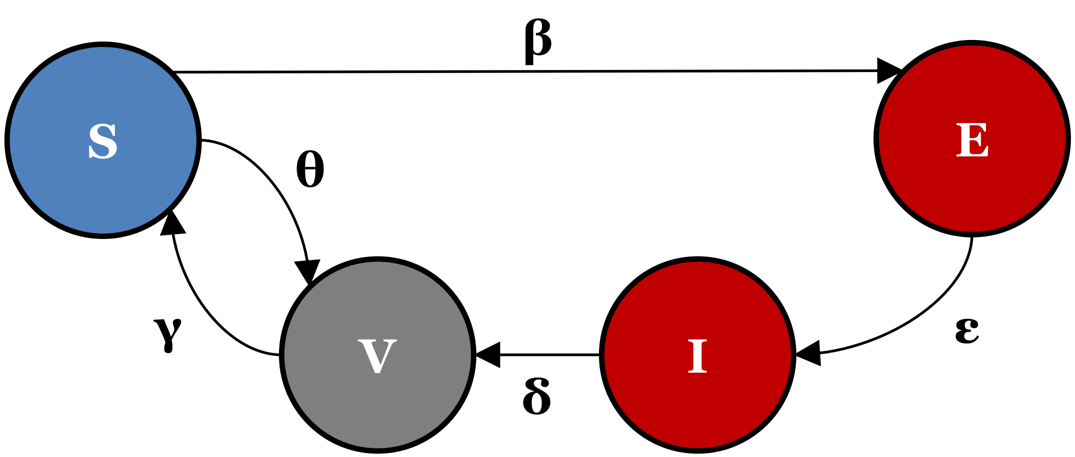
SIRS VPM

- ***SIRS* VPM** ("susceptible, infected, recovered, susceptible"):
  - Example: pertussis.
  - Three states: **Susceptible** (S), **Infected** (I), **Recovered** (R).
  - Transition probabilities: attack/transmission probability β, healing probability δ, immunization loss probability γ.
  - Temporary immunity.



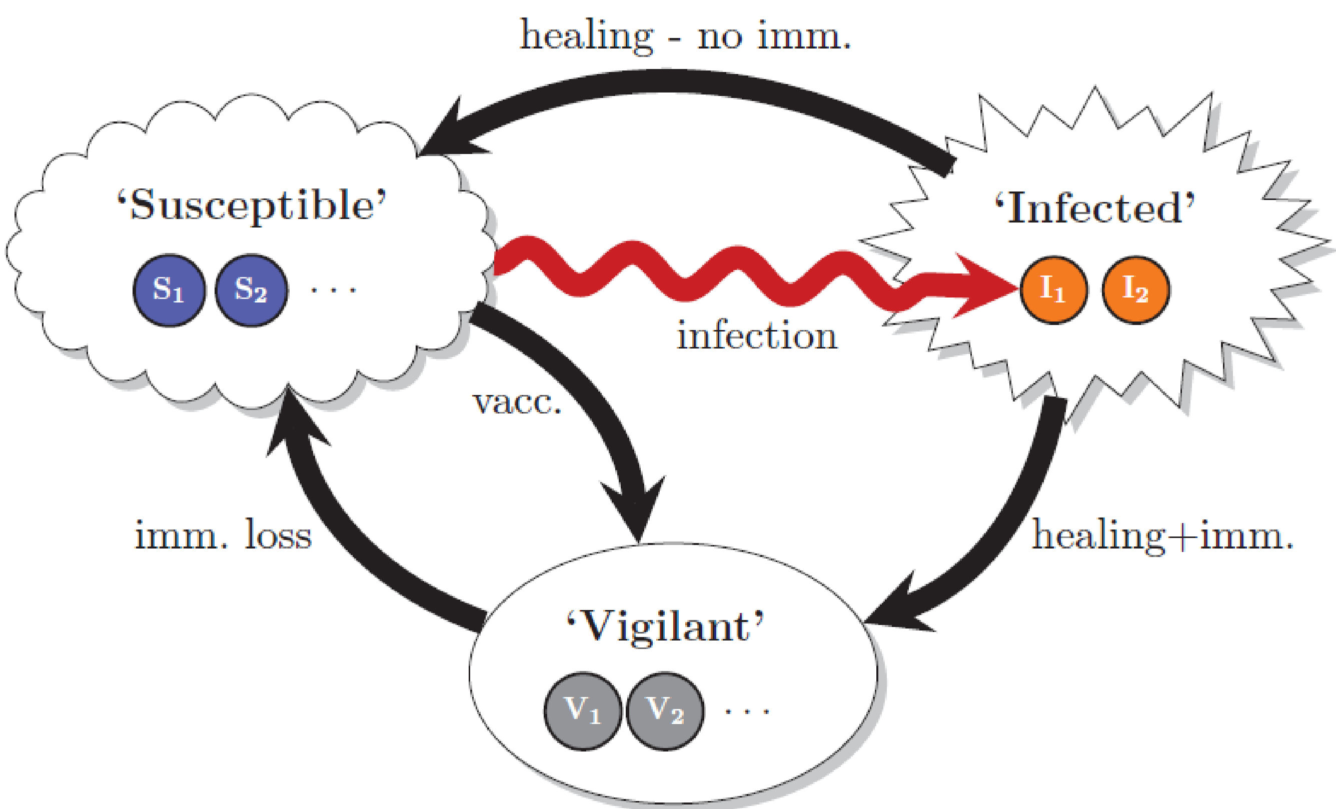
SEIV **VPM**

- ***SEIV*** **VPM** (“susceptible, exposed, infected, vigilant”):
  - States: **Susceptible** (S), **Exposed** (E), **Infected** (I), **Vigilant** (V).
  - Transition probabilities: attack/transmission probability β, healing probability δ, immunization loss probability γ, virus maturation probability ϵ, direct immunization (vaccination) probability θ.
  - Temporary immunity and virus incubation.



$S^* I^2 V^*$ VPM

- All existing VPMs can be generalized to the $S^* I^2 V^*$ **VPM**.



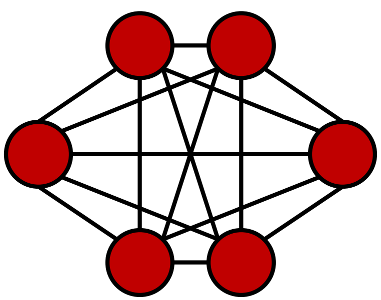
Source: Chakrabarti and Faloutsos, 2012

Epidemic Threshold

- The **epidemic threshold** captures the transition in the behavior of the system:
  - Above the threshold, the virus can spread across the network and result in a network-wide epidemic.
  - Below the threshold, the virus can't spread across the network and will die quickly.
- The **epidemic threshold** depends on:
  - The characteristics of the VPM.
  - The **connectivity** of the underlying contact network.

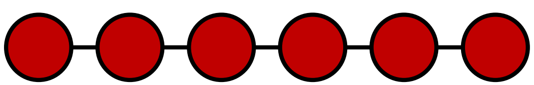
Contact Network Connectivity

- A better connected contact network facilitates the spread of the virus.
- How to measure the **connectivity** of the contact network?
  - Average degree?



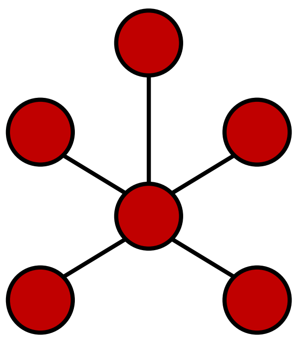
Clique

Avg. degree = 5



Chain

Avg. degree = 1.67



Star

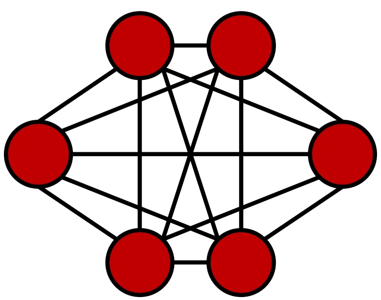
Avg. degree = 1.67

Oak Ridge National Laboratory

NC STATE UNIVERSITY Department of Computer Science

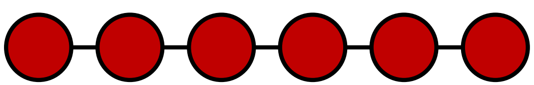
Contact Network Connectivity

- The **spectral radius** (i.e., **largest eigenvalue**) of the adjacency matrix of the network (λ_1) measures the connectivity of the graph better than the average degree because it takes into account all path-lengths.



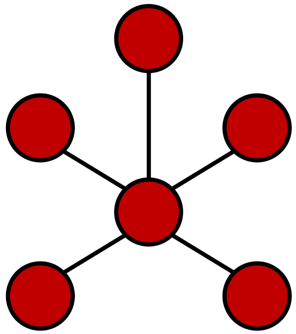
Clique

$\lambda_1 = n - 1 = 5$



Chain

$\lambda_1 = 1.80$



Star

$\lambda_1 = \sqrt{n-1} = \sqrt{5} = 2.24$

Effective Strength

- **Theorem** [*Prakash et al., 2010*]:
  - For any VPM that follows the $S^*I^2V^*$ model and for any arbitrary contact network with adjacency matrix A, the **effective strength** s of a virus is:

$$s = \lambda_1 \cdot C_{VPM}$$

    where λ_1 is the largest eigenvalue of A and C_{VPM} is a constant that depends on the VPM.
  - The **epidemic threshold** is reached when:

$$s = 1$$

  - The **sufficient condition for stability** is given by:

$$s < 1$$

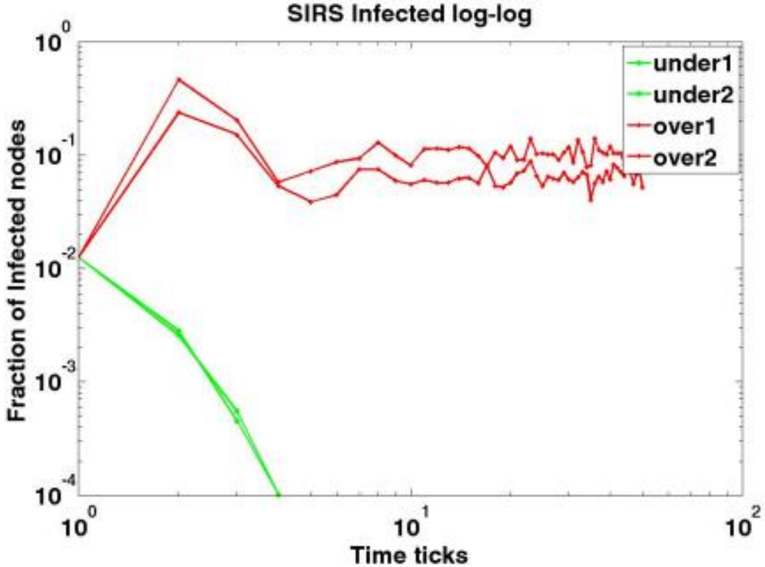
Effective Strength

- Effective strength for some VPMs:

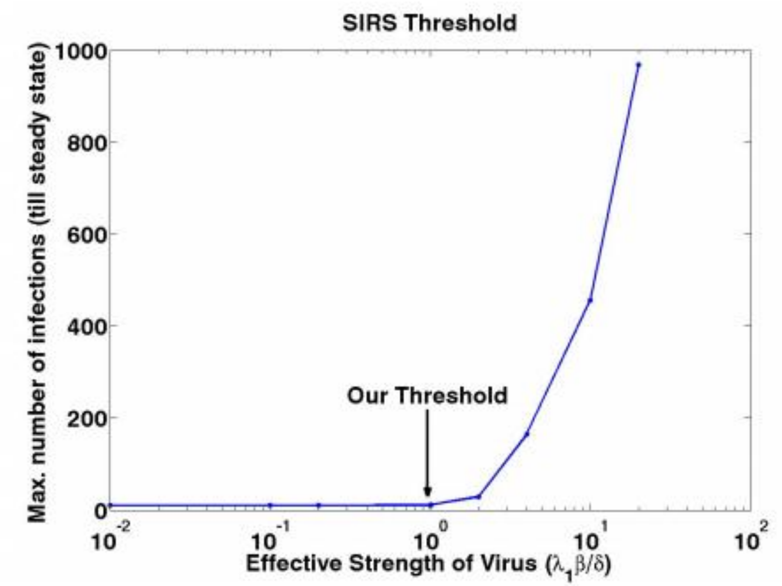
VPM	C_{VPM}	Effective Strength
SIS	$\frac{\beta}{\delta}$	$\lambda_1 \cdot \frac{\beta}{\delta}$
SIR	$\frac{\beta}{\delta}$	$\lambda_1 \cdot \frac{\beta}{\delta}$
$SIRS$	$\frac{\beta}{\delta}$	$\lambda_1 \cdot \frac{\beta}{\delta}$
$SEIV$	$\frac{\beta\gamma}{\delta(\gamma+\theta)}$	$\lambda_1 \cdot \frac{\beta\gamma}{\delta(\gamma+\theta)}$

Effective Strength

- Results of simulation experiments on Oregon autonomous system (AS) router graph (3,995 nodes, 15,420 edges):



(A) SIRS Infective Fraction Time Plot (log-log)



(B) SIRS Max. Infections till steady state vs Strength (lin-log)

Source: Prakash et al., 2010

Alternating Contact Networks

- How can we extend this theorem to time-varying graphs with **alternating connectivity behavior**?
- **Theorem** [*Prakash et al., 2010*]:
  - Given a set of T adjacency matrices $\{A_1, A_2, \dots, A_T\}$ for alternating contact networks and the SIS VPM with parameters β and δ, the **effective strength** s of a virus is:

    $$s = \lambda_{\prod_{i=1}^{T} S_i}$$

    where $\prod_{i=1}^{T} S_i$ is called the **system-matrix**,

    λ is the largest eigenvalue of the **system-matrix**,

    and $S_i = (1-\delta)I + \beta A_i \ \forall i \in \{1,2,\dots,T\}$.

Immunization Policies

- Given a contact network and a number of available vaccines k, an **immunization policy** determines which are the k best nodes to immunize (i.e., remove from contact network) in order to prevent the virus from spreading across the network.
- Some immunization policies:
  - Randomly selecting k nodes for immunization.
  - Selecting k **high-degree nodes** for immunization.
  - Selecting k nodes whose removal will cause the largest drop in the value of λ_1 for immunization (**largest eigen-drop**).

OAK RIDGE National Laboratory

NC STATE UNIVERSITY Department of Computer Science

Optimal Immunization Policy

- The optimal immunization policy is to find the subset of k nodes with the largest **eigen-drop** among all $\binom{n}{k}$ possible subsets.
  - This solution is computationally intractable due to its combinatorial nature.
  - If it takes about 0.01 seconds to compute λ_1, it would take 2,615 years to find the best subset of 5 nodes in a graph with 1,000 nodes and 10,000 edges.

Oak Ridge National Laboratory

NC STATE UNIVERSITY
Department of Computer Science

Heuristic Immunization Policy

- The **eigen-drop** caused by a set of nodes S can be approximated by calculating the ***Shield-value*** score [*Tong et al., 2010*]:

$$Sv(S) = \sum_{i \in S} 2\lambda_1 u_1(i)^2 - \sum_{i,j,\in S} A(i,j) u_1(i) u_1(j)$$

where A is the adjacency matrix of the network, λ_1 is the largest eigenvalue of A and u_1 is the corresponding eigenvector.

  – Finding the subset of k nodes with the highest ***Shield-value*** score among all $\binom{n}{k}$ possible subsets is still infeasible for most real-world applications.

NetShield Algorithm

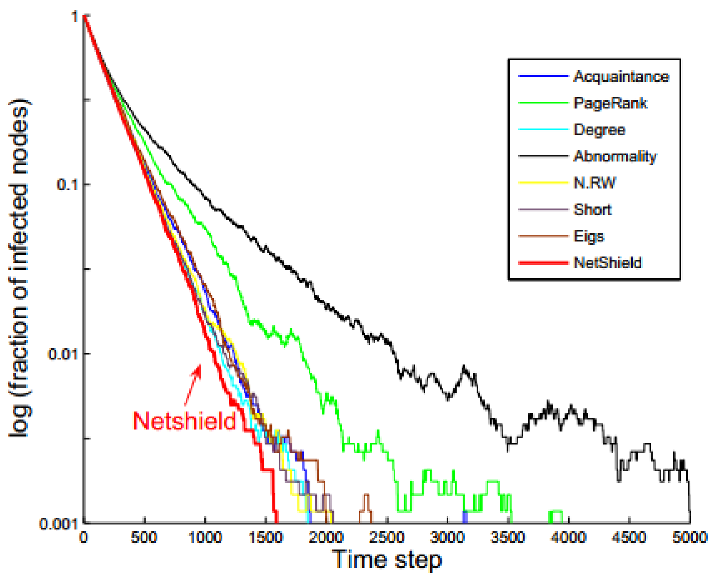
[*Tong et al., 2010*]

1. Compute the largest eigenvalue λ_1 and the corresponding eigenvector u_1 of the contact network's adjacency matrix A.
2. For each node i in the contact network, calculate the ***Shield-value*** score $Sv(i)$.
3. Initialize an empty subset S.
4. Let $\mathrm{b} = A(:,S) \cdot u_1(S)$.
5. For each node i in the contact network, compute:
$$score(i) = Sv(i) - 2 \cdot b(i) \cdot u_1(i)$$
6. Add the node i with the maximum $score(i)$ to the subset S.
7. Repeat (4)-(6) until S contains k nodes.

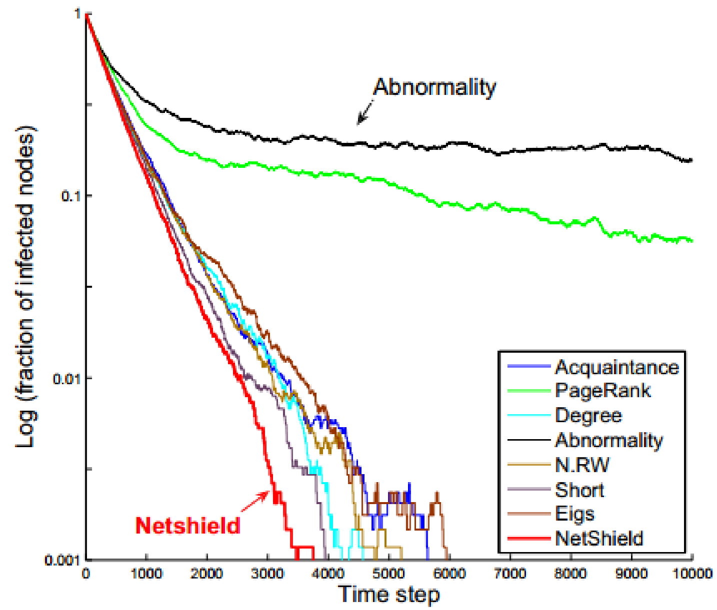
NetShield Algorithm

[*Tong et al., 2010*]

- Evaluation of the **NetShield Algorithm** and other immunization policies on the Karate graph (34 nodes, 152 edges):



(a) $s = 1.4$



(b) $s = 2.9$

Source: Tong et al., 2010

Bonus Problem: Viral Marketing

- You are in charge of a **viral marketing** campaign to introduce a new product. Your goal is to disseminate information about this product across a social network.
- Your budget allows you to target only k individuals by directly sharing information about the product. You expect these individuals to then share this information with their contacts.
- What would be your strategy to maximize the spread of the information about the product across the social network?

References

- D. Chakrabarti, C. Faloutsos. *Graph Mining: Laws, Tools, and Case Studies*. Synthesis Lectures on Data Mining and Knowledge Discovery, Morgan & Claypool Publishers, 2012
- B. A. Prakash, D. Chakrabarti, M. Faloutsos, N. Valler, C. Faloutsos. Got the Flu (or Mumps)? Check the Eigenvalue! *arXiv:1004.0060 [physics.soc-ph]*, 2010.
- B. A. Prakash, H. Tong, N. Valler, M. Faloutsos, C. Faloutsos. Virus Propagation on Time-Varying Networks: Theory and Immunization Algorithms. In *ECML-PKDD*, 2010.
- H. Tong, B. A. Prakash, C. Tsourakakis, T. Eliassi-Rad, C. Faloutsos, D. H. Chau. On the Vulnerability of Large Graphs. In *ICDM*, 2010.

Thank You!