Understanding and Managing Cascades on Large Graphs

B. Aditya Prakash

Virginia Tech.

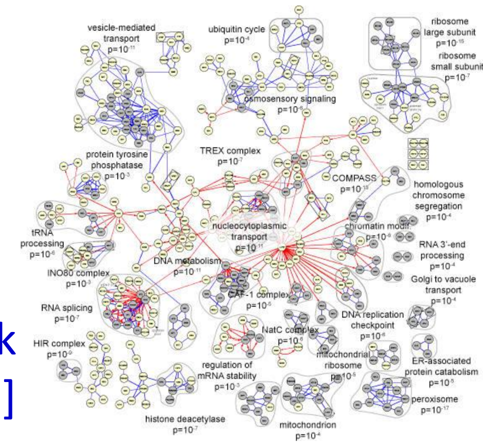
Christos Faloutsos

Carnegie Mellon University

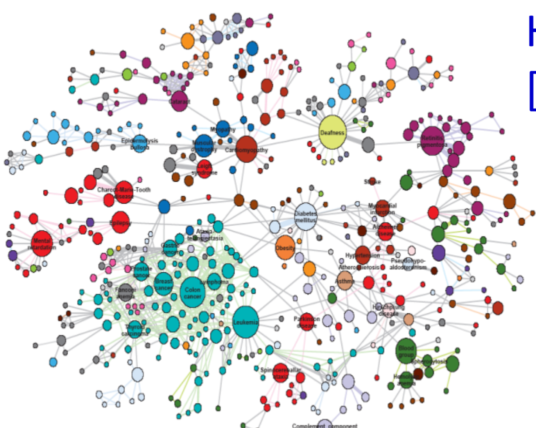
Sept 28, Tutorial, ECML-PKDD 2012, Bristol

Networks are everywhere!

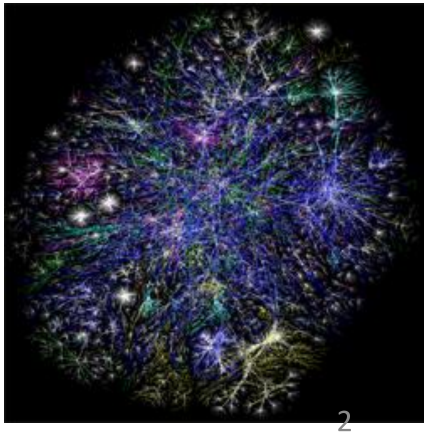
Facebook Network [2010]



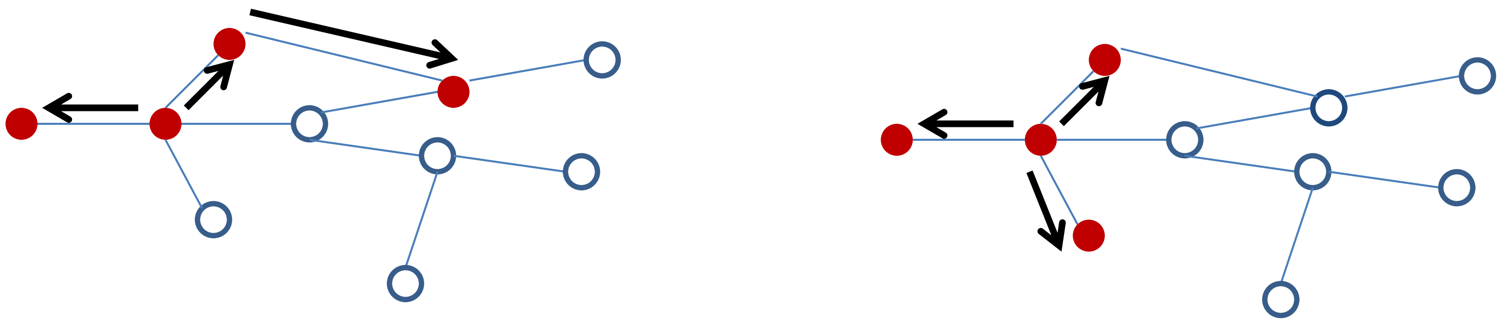
Gene Regulatory Network [Decourty 2008]



Human Disease Network [Barabasi 2007]



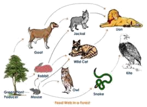
The Internet [2005]



Dynamical Processes ***over*** networks are also everywhere!

Why do we care?

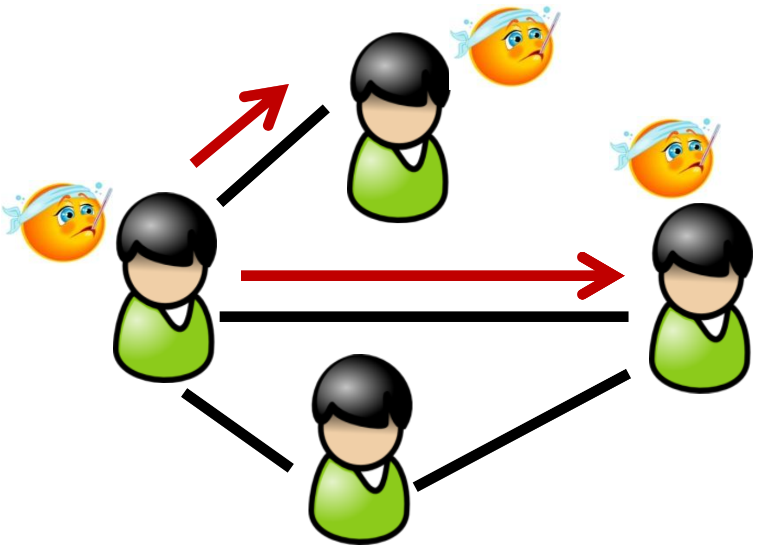
- Social collaboration
- Information Diffusion
- Viral Marketing
- Epidemiology and Public Health
- Cyber Security
- Human mobility
- Games and Virtual Worlds
- Ecology



....

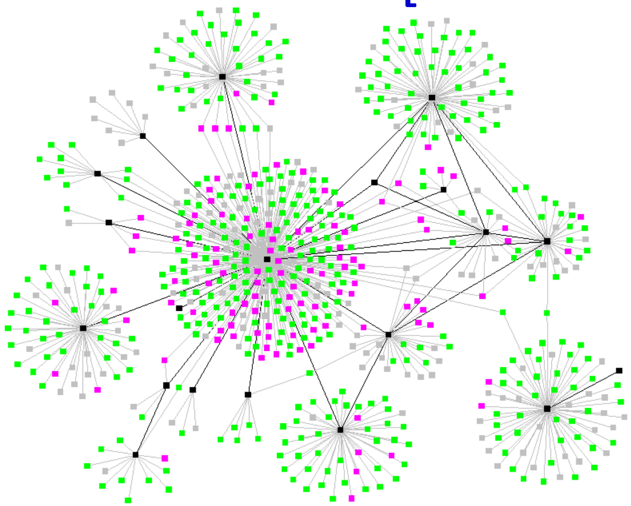
Why do we care? (1: Epidemiology)

- Dynamical Processes over networks



Diseases over contact networks

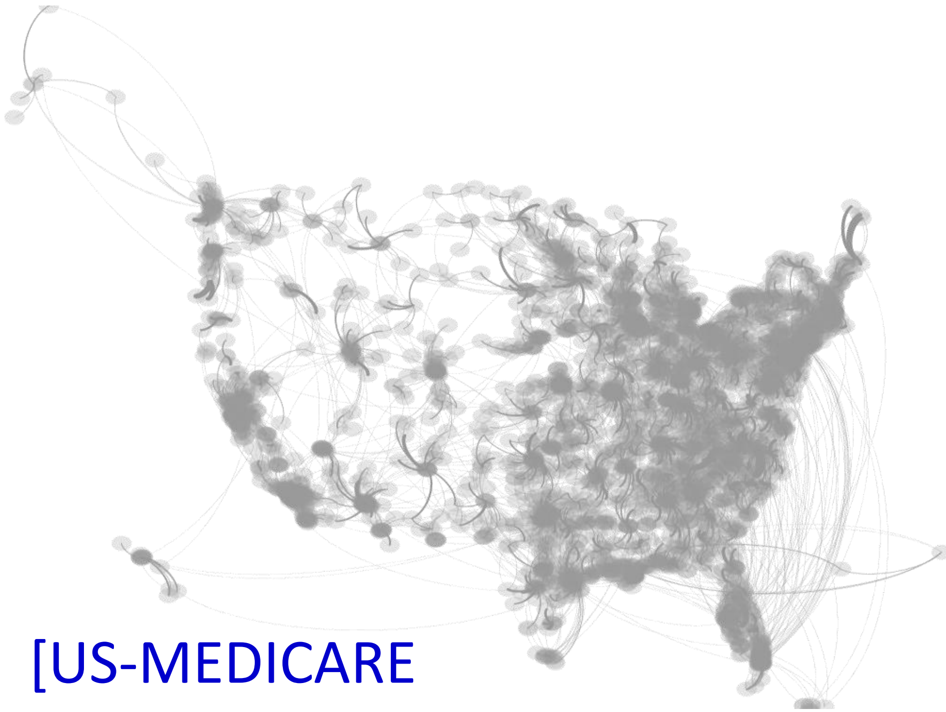
[AJPH 2007]



CDC data: Visualization of the first 35 tuberculosis (TB) patients and their 1039 contacts

Why do we care? (1: Epidemiology)

- Dynamical Processes over networks



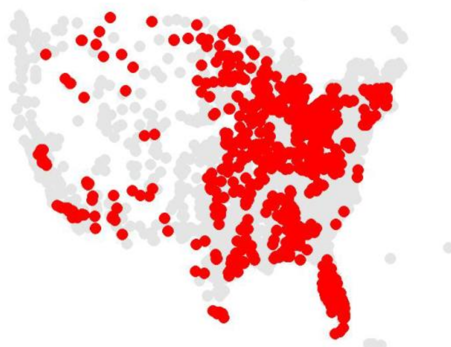
- Each circle is a hospital
- ~3000 hospitals
- More than 30,000 patients transferred

[US-MEDICARE NETWORK 2005]

Problem: Given *k* units of disinfectant, whom to immunize?

Why do we care? (1: Epidemiology)

~6x fewer!



CURRENT PRACTICE

[US-MEDICARE NETWORK 2005]

OUR METHOD

Hospital-acquired inf. took 99K+ lives, cost $5B+ (all per year)

VirginiaTech

Carnegie Mellon

Why do we care? (2: Online Diffusion)

licensed under (cc) Attribution-NonCommercial-ShareAlike 2.0 Germany | Ludwig Gatzke | http://flickr.com/photos/stabilo-boss/

> 800m users, ~$1B revenue [WSJ 2010]

facebook

~100m active users

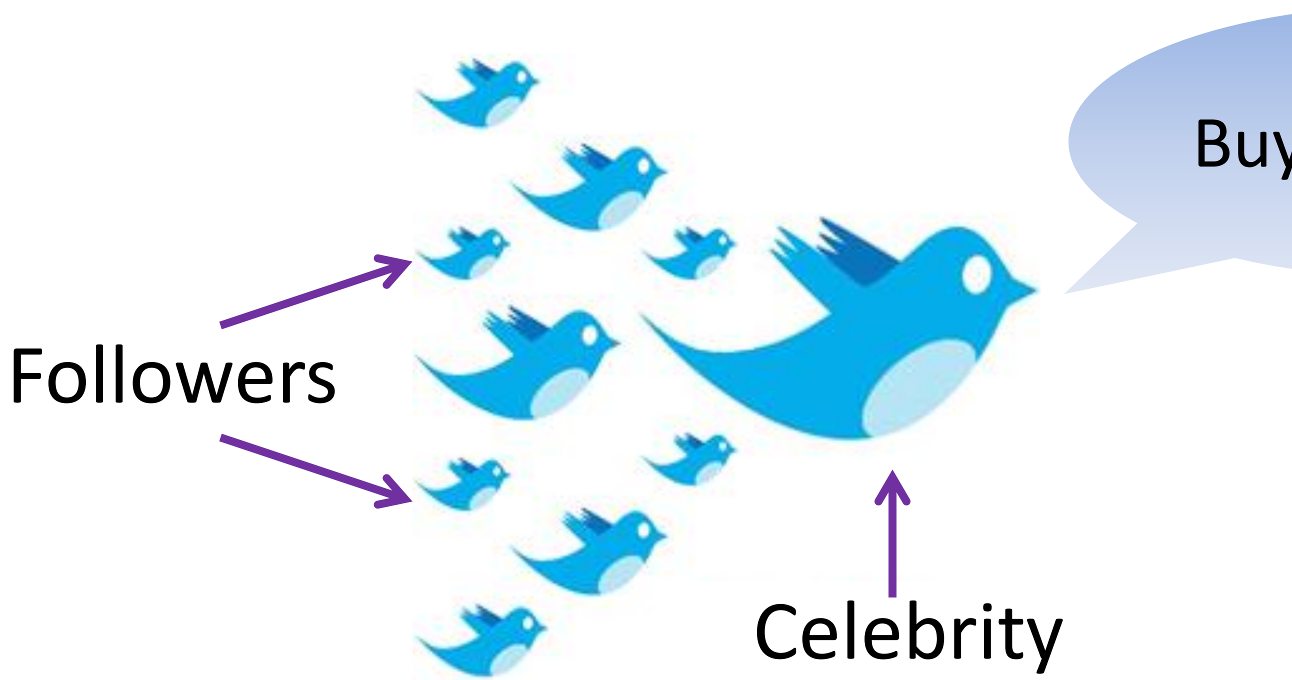
twitter

> 50m users

LinkedIn

Why do we care? (2: Online Diffusion)

- Dynamical Processes over networks



Social Media Marketing

Why do we care? (3: To change the world?)

- Dynamical Processes over networks

Social networks and Collaborative Action

High Impact – Multiple Settings

*Q. How to squash **rumors** faster?*

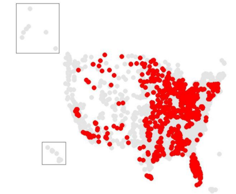
*Q. How do **opinions** spread?*

*Q. How to **market** better?*

High Impact – Multiple Settings

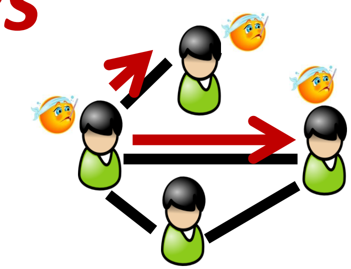
epidemic out-breaks

Q. How to squash ~~***rumors***~~ *faster?*



products/viruses

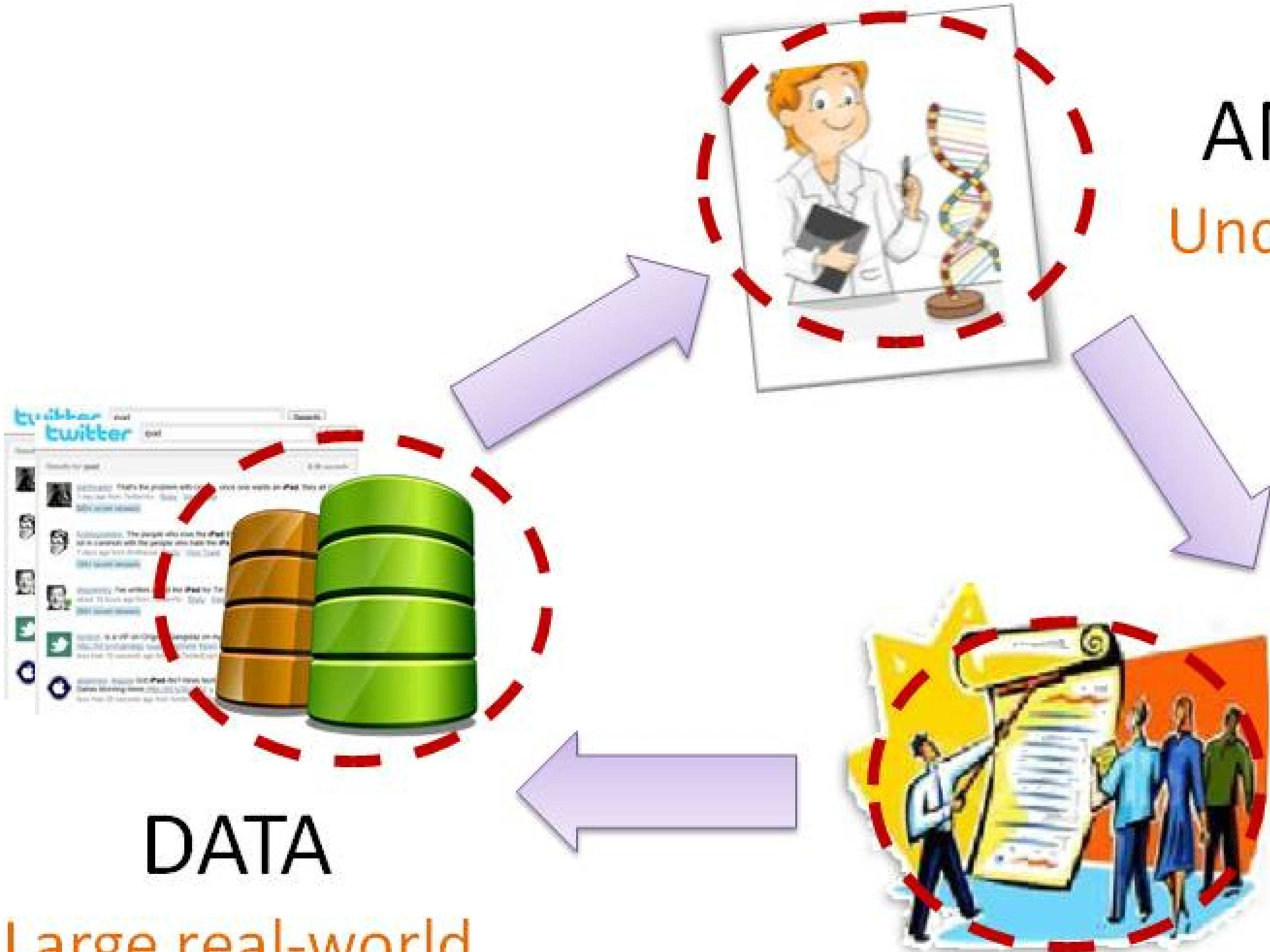
Q. How do ~~***opinions***~~ *spread?*



transmit s/w patches

Q. How to ~~***market***~~ *better?*

Research Theme



ANALYSIS

Understanding

POLICY/ ACTION

Managing

DATA

Large real-world networks & processes

Research Theme – Public Health

ANALYSIS

Will an epidemic happen?

DATA

Modeling # patient transfers

POLICY/ ACTION

How to control out-breaks?

Research Theme – Social Media

ANALYSIS

cascades in future?

DATA

Modeling Tweets spreading

POLICY/ ACTION

How to market better?

In this tutorial

ANALYSIS

Understanding

Given propagation models:

Q1: What is the epidemic threshold?

Q2: How do viruses compete?

In this tutorial

POLICY/
ACTION
Managing

Q3: How to immunize and control out-breaks better?

Q4: How to detect outbreaks?

Q5: Who are the culprits?

In this tutorial

DATA

Large real-world networks & processes

Q6: How do cascades look like?

Q7: How does activity evolve over time?

Q8: How does external influence act?

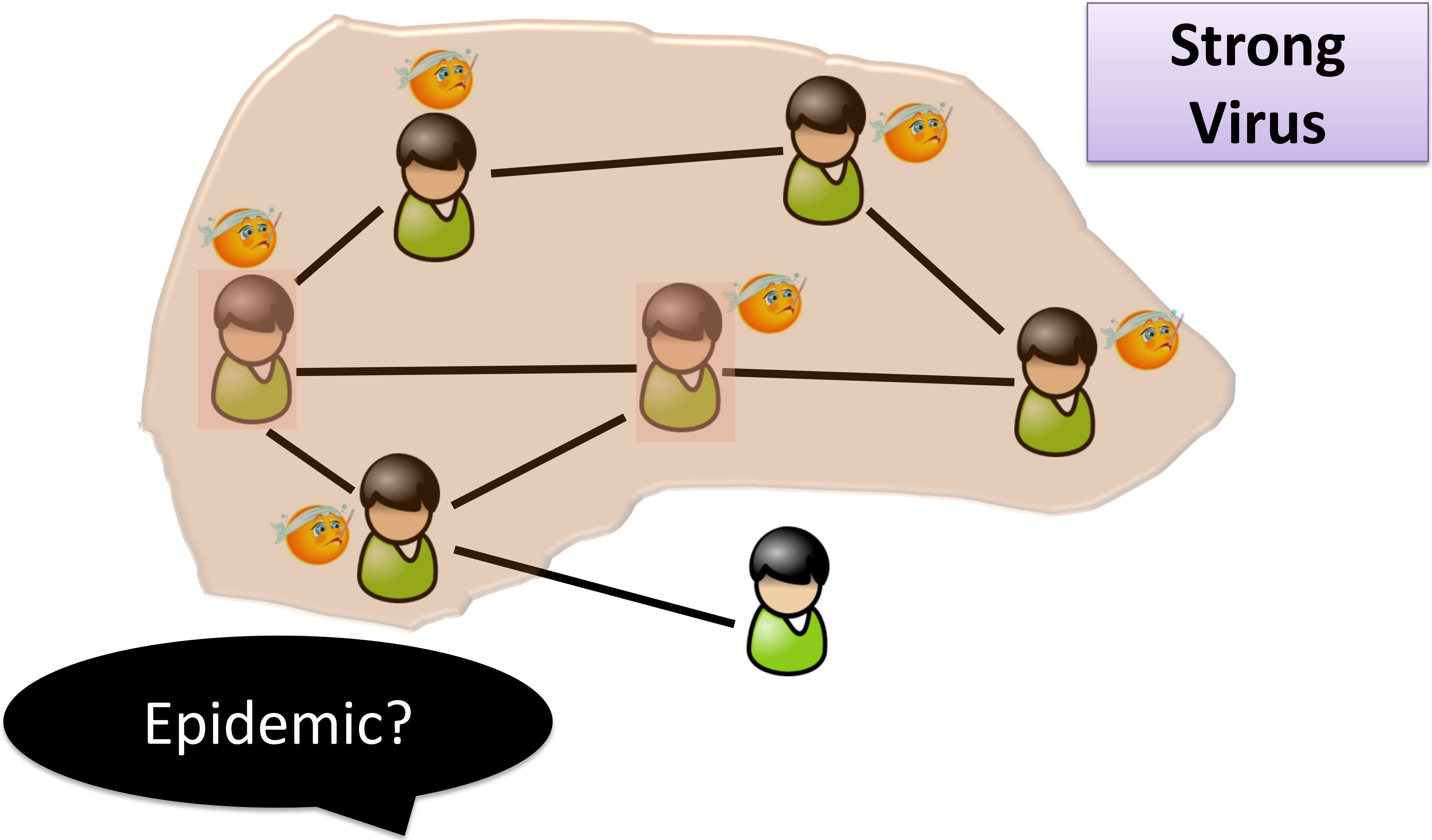
Outline

- Motivation
- **Part 1: Understanding Epidemics (Theory)**
- Part 2: Policy and Action (Algorithms)
- Part 3: Learning Models (Empirical Studies)
- Conclusion

Part 1: Theory

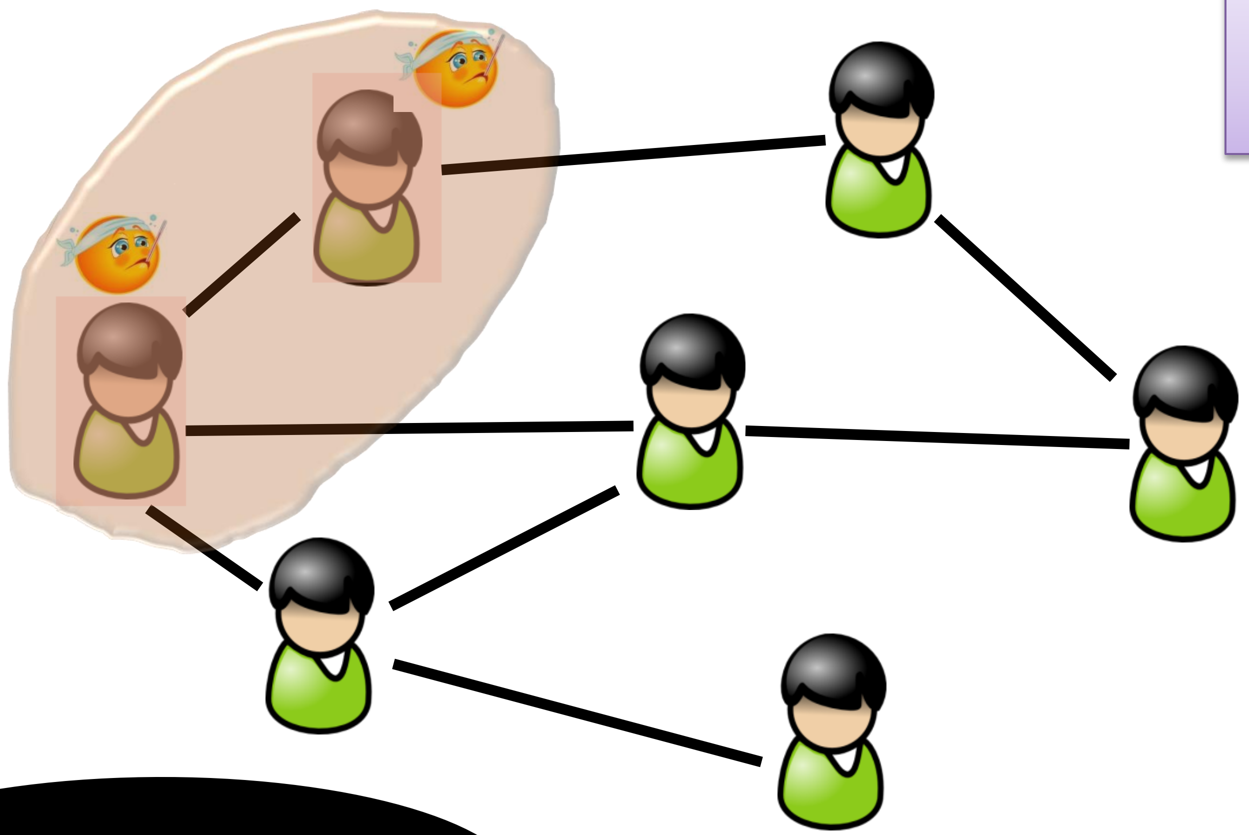
- **Q1: What is the epidemic threshold?**
- Q2: How do viruses compete?

A fundamental question



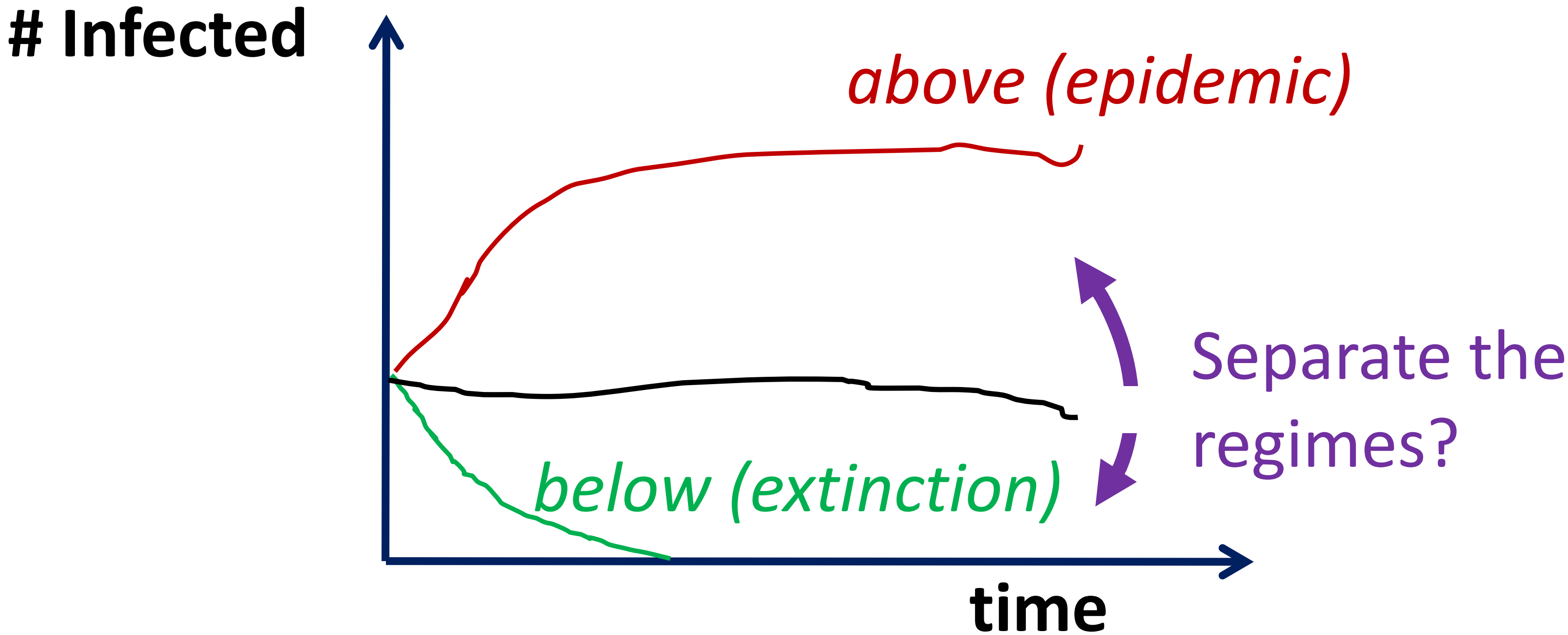
example (static graph)

Weak Virus



Epidemic?

Problem Statement



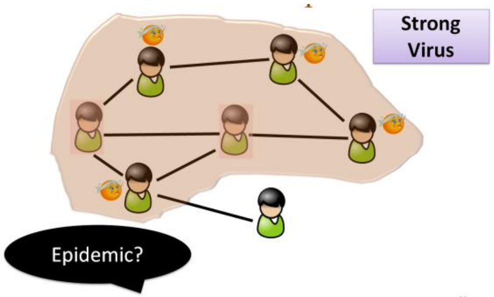
Find, a condition under which

- *virus will die out exponentially quickly*
- *regardless of initial infection condition*

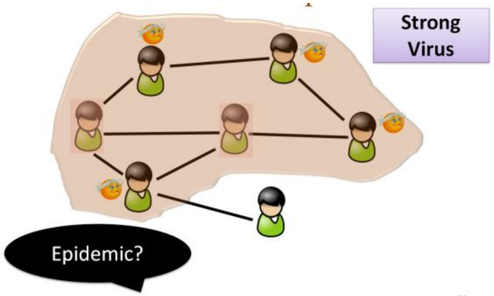
Threshold (static version)

Problem Statement

- Given:
  - Graph G, *and*
  - Virus specs (attack prob. etc.)
- Find:
  - A condition for virus extinction/invasion

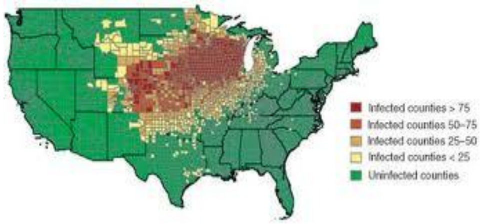


Threshold: Why important?



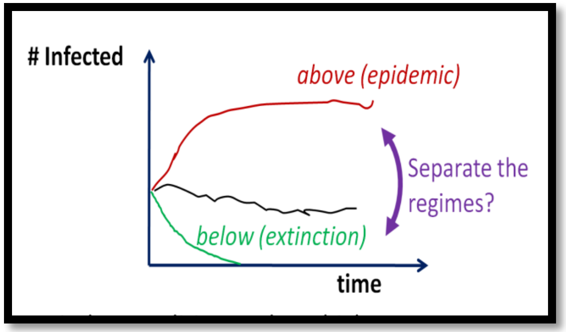
- Accelerating simulations
- Forecasting ('What-if' scenarios)
- Design of contagion and/or topology
- A great handle to manipulate the spreading
  - Immunization
  - Maximize collaboration

  .....



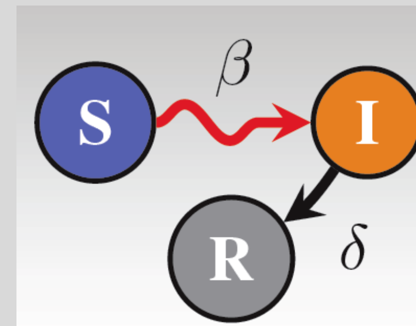
Part 1: Theory

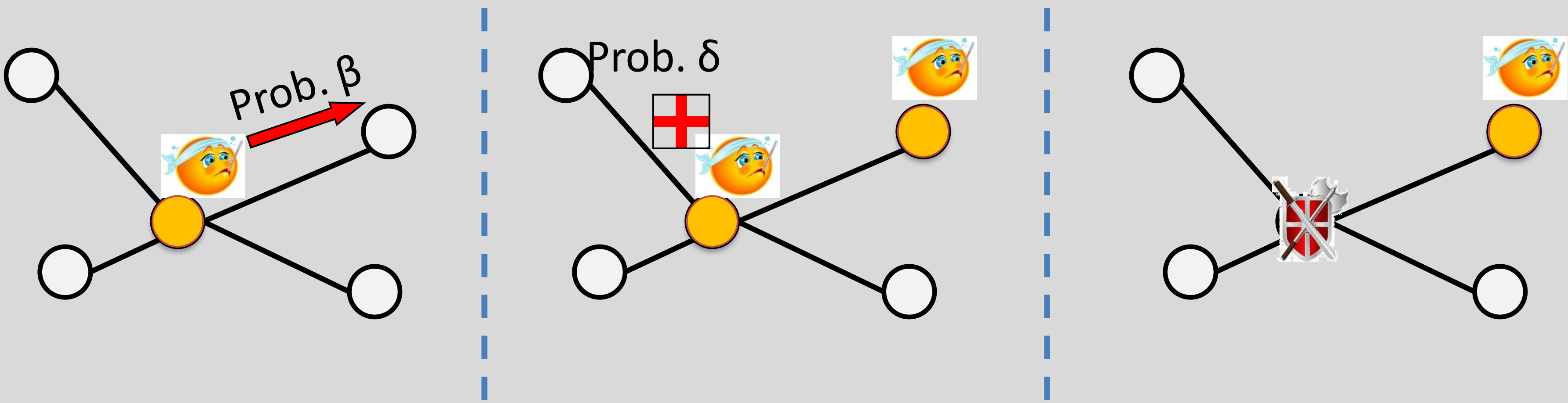
- Q1: What is the epidemic threshold?
  - **Background**
  - Result and Intuition (Static Graphs)
  - Proof Ideas (Static Graphs)
  - Bonus: Dynamic Graphs
- Q2: How do viruses compete?



"SIR" model: life immunity (mumps)

- Each node in the graph is in one of three states
  - **S**usceptible (i.e. healthy)
  - **I**nfected
  - **R**emoved (i.e. can't get infected again)





t = 1 *t = 2* *t = 3*

Terminology: continued

- Other virus propagation models ("VPM")
  - SIS : susceptible-infected-susceptible, flu-like
  - SIRS : temporary immunity, like pertussis
  - SEIR : mumps-like, with virus incubation
    (E = Exposed)

  ……………….
- Underlying contact-network – 'who-can-infect-whom'

Related Work

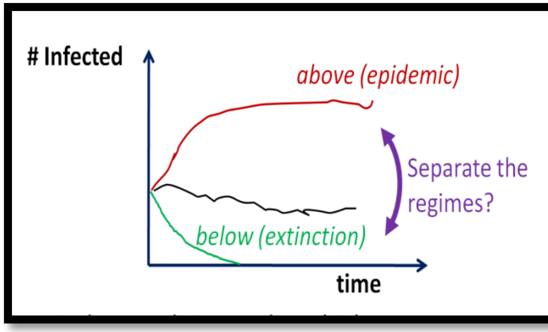
- R. M. Anderson and R. M. May. Infectious Diseases of Humans. Oxford University Press, 1991.
- A. Barrat, M. Barthélemy, and A. Vespignani. Dynamical Processes on Complex Networks. Cambridge University Press, 2010.
- F. M. Bass. A new product growth for model consumer durables. Management Science, 15(5):215–227, 1969.
- D. Chakrabarti, Y. Wang, C. Wang, J. Leskovec, and C. Faloutsos. Epidemic thresholds in real networks. ACM TISSEC, 10(4), 2008.
- D. Easley and J. Kleinberg. Networks, Crowds, and Markets: Reasoning About a Highly Connected World. Cambridge University Press, 2010.
- A. Ganesh, L. Massoulie, and D. Towsley. The effect of network topology in spread of epidemics. IEEE INFOCOM, 2005.
- Y. Hayashi, M. Minoura, and J. Matsukubo. Recoverable prevalence in growing scale-free networks and the effective immunization. arXiv:cond-at/0305549 v2, Aug. 6 2003.
- H. W. Hethcote. The mathematics of infectious diseases. SIAM Review, 42, 2000.
- H. W. Hethcote and J. A. Yorke. Gonorrhea transmission dynamics and control. Springer Lecture Notes in Biomathematics, 46, 1984.
- J. O. Kephart and S. R. White. Directed-graph epidemiological models of computer viruses. IEEE Computer Society Symposium on Research in Security and Privacy, 1991.
- J. O. Kephart and S. R. White. Measuring and modeling computer virus prevalence. IEEE Computer Society Symposium on Research in Security and Privacy, 1993.
- R. Pastor-Santorras and A. Vespignani. Epidemic spreading in scale-free networks. Physical Review Letters 86, 14, 2001.
- ………
- ………
- ………

All are about *either*:

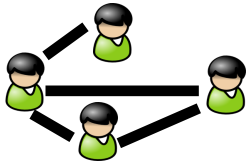
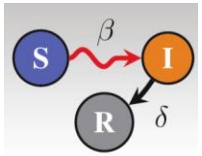
- **Structured topologies** (cliques, block-diagonals, hierarchies, random)
- **Specific virus propagation models**
- **Static graphs**

Part 1: Theory

- Q1: What is the epidemic threshold?
  - Background
  - **Result and Intuition (Static Graphs)**
  - Proof Ideas (Static Graphs)
  - Bonus: Dynamic Graphs
- Q2: How do viruses compete?



How should the answer look like?

- Answer should depend on:
  - Graph 
  - Virus Propagation Model (VPM) 

- But how??
  - Graph – average degree? max. degree? diameter?
  - VPM – which parameters?
  - How to combine – linear? quadratic? exponential?

$$\beta d_{avg} + \delta\sqrt{diameter} \,?\quad (\beta^2 d^2{}_{avg} - \delta d_{avg}) / d_{\max} \,?\quad .....$$

Static Graphs: Our Main Result

- Informally,

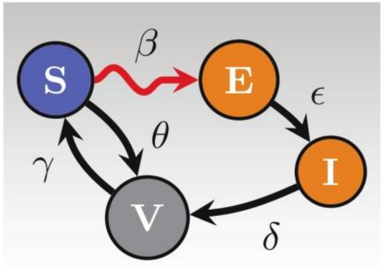
For,
- any arbitrary topology (adjacency matrix A)
- any virus propagation model (VPM) in standard literature

the epidemic threshold depends only
1. on the λ, first eigenvalue of A, *and*
2. some constant C_{VPM} , determined by the virus propagation model

any arbitrary topology (adjacency matrix A)	**λ**
any virus propagation model (VPM) in standard literature	C_{VPM}
the epidemic threshold depends only on λ and C_{VPM}	**No epidemic if $\lambda * C_{VPM} < 1$**

In Prakash+ ICDM 2011

Our thresholds for some models



- *s = effective strength*
- *s* < 1 : *below threshold*

Models	Effective Strength (s)	Threshold (tipping point)
SIS, SIR, SIRS, SEIR	s = λ . $\left(\frac{\beta}{\delta}\right)$	s = 1
SIV, SEIV	s = λ . $\left(\frac{\beta\gamma}{\delta(\gamma+\theta)}\right)$	
$SI_1I_2V_1V_2$ (H.I.V.)	s = λ . $\left(\frac{\beta_1 v_2 + \beta_2 \varepsilon}{v_2(\varepsilon + v_1)}\right)$	

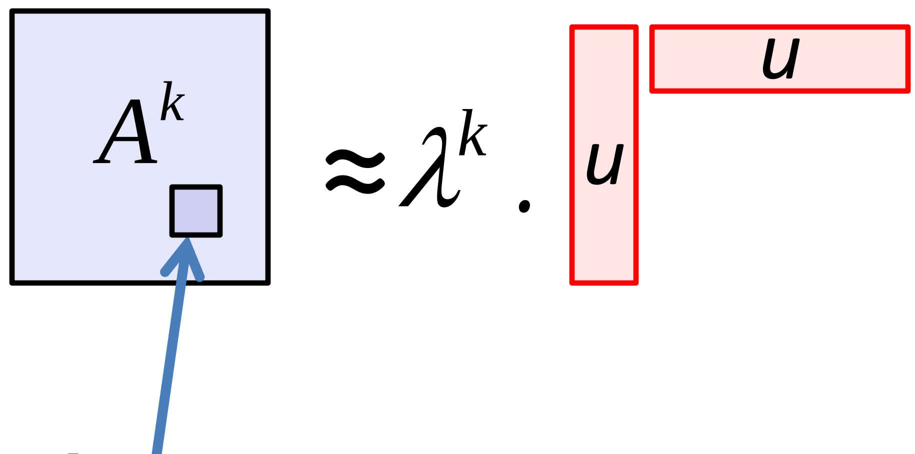
Our result: Intuition for λ

"Official" definition:

- *Let A be the adjacency matrix. Then λ is the root with the largest magnitude of the characteristic polynomial of A [$\det(A - xI)$].*
- Doesn't give much intuition!

"Un-official" Intuition ☺

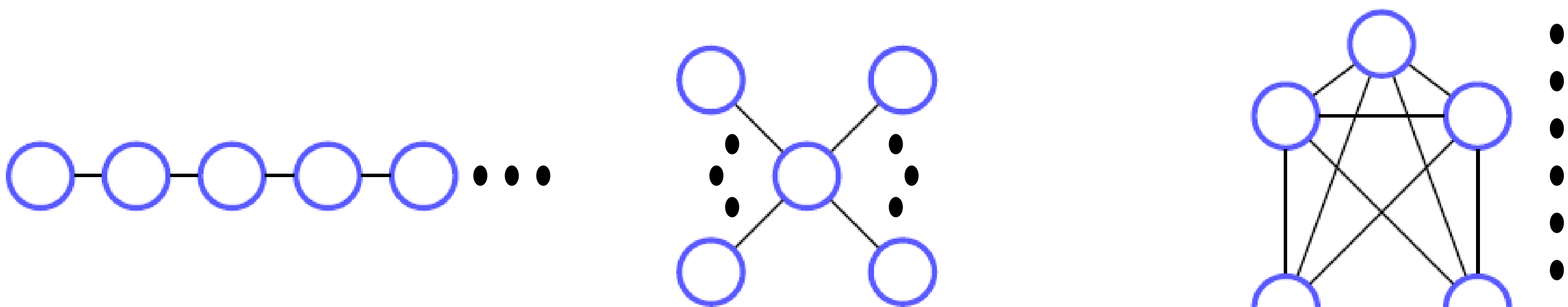
- λ ~ # paths in the graph



A^k(i, j) = # of paths i → j of length k

Largest Eigenvalue (λ)

better connectivity $\longrightarrow$ higher λ



$\lambda \approx 2$ $\lambda = \sqrt{N}$ $\lambda = N\text{-}1$

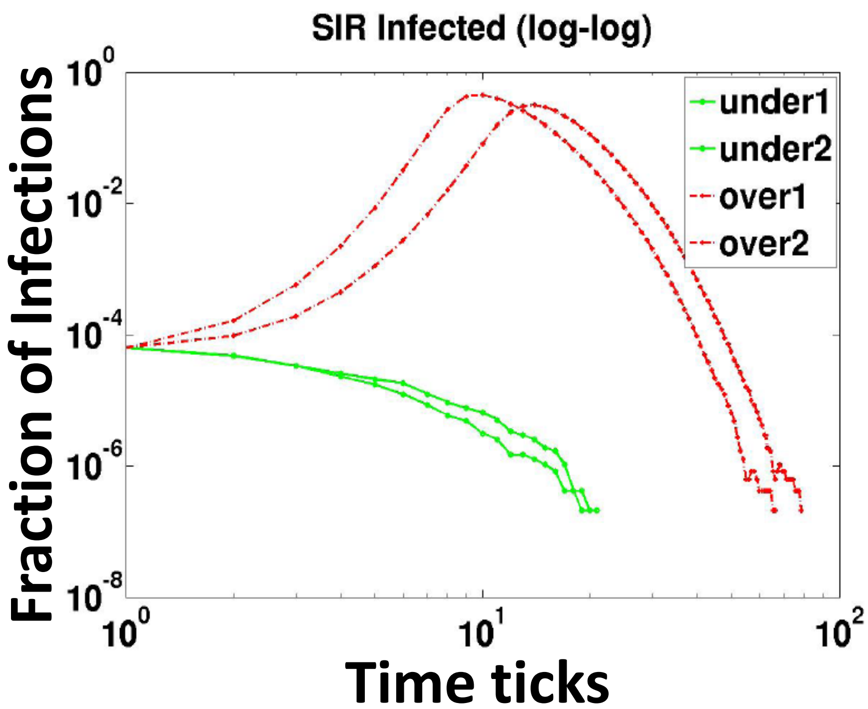
(a)Chain (b)Star (c)Clique

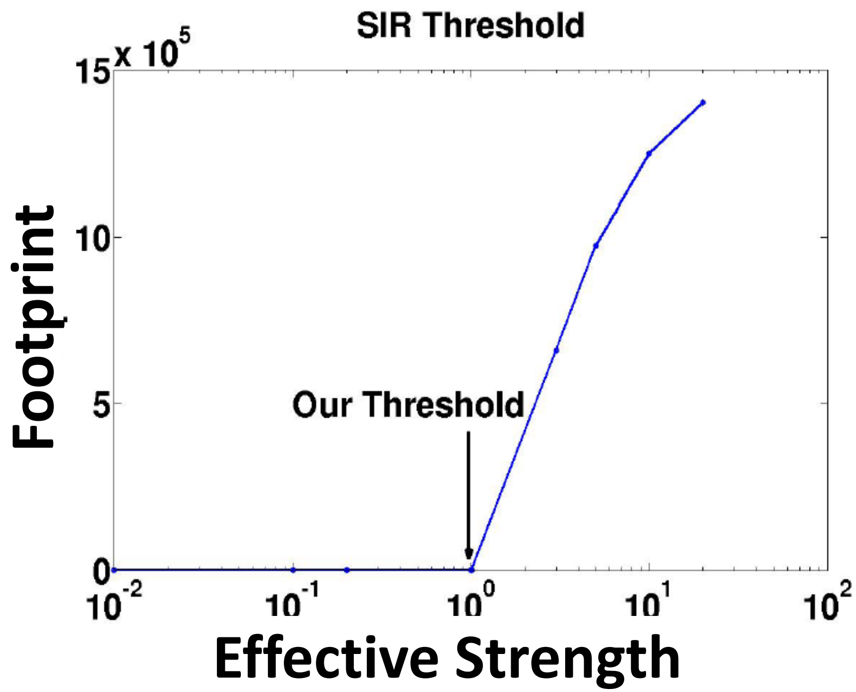
$\lambda \approx 2$ $\lambda = 31.67$ $\lambda = 999$

$N = 1000$

N nodes

Examples: Simulations – SIR (mumps)



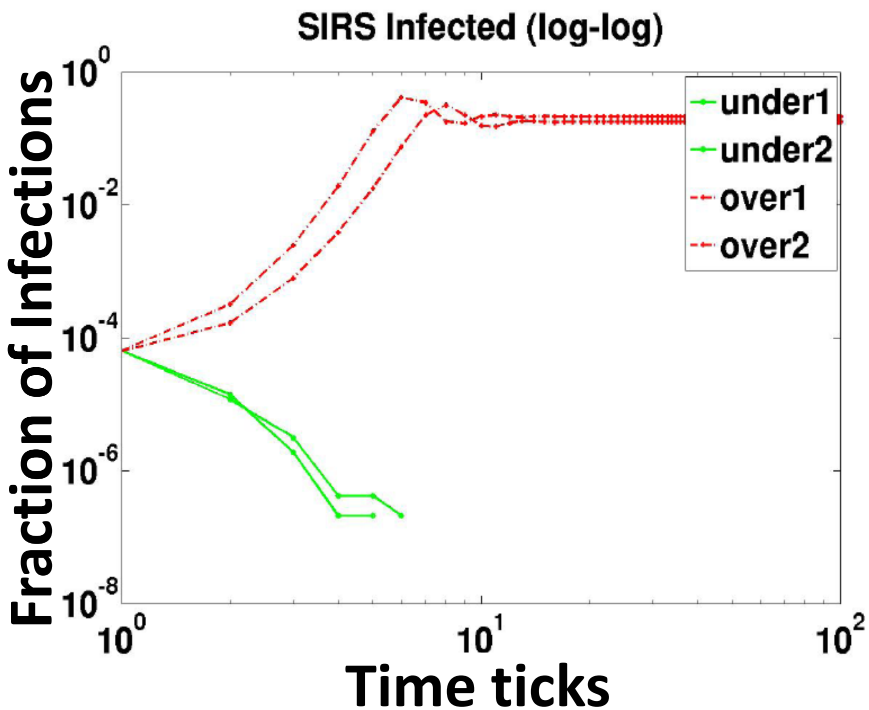


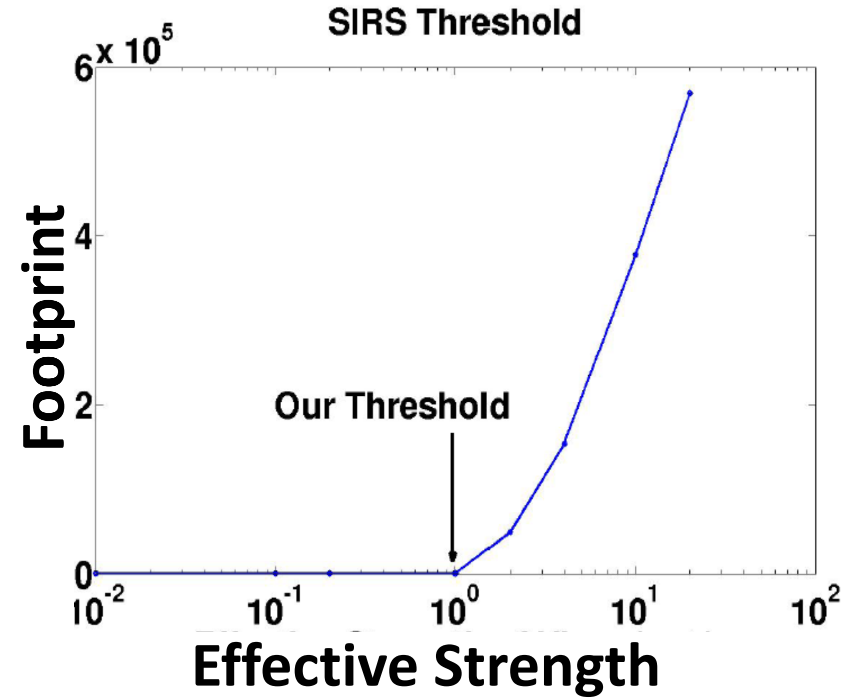
(a) Infection profile **(b) "Take-off" plot**

PORTLAND graph

31 million links, 6 million nodes

Examples: Simulations – SIRS (pertusis)





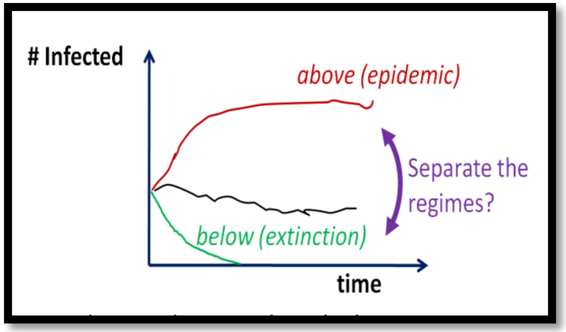
(a) Infection profile **(b) "Take-off" plot**

PORTLAND graph

31 million links, 6 million nodes

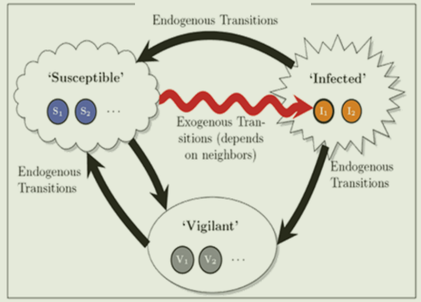
Part 1: Theory

- Q1: What is the epidemic threshold?
  - Background
  - Result and Intuition (Static Graphs)
  - **Proof Ideas (Static Graphs)**
  - Bonus: Dynamic Graphs
- Q2: How do viruses compete?



Proof Sketch

Model	Used for
SIR	Mumps
SIS	Flu
SIRS	Pertussis
SEIR	Varicella
……..	
SICR	Tuberculosis
MSIR	Measles
SIV	Sensor Stability
$SI_1I_2V_1V_2$	H.I.V.
……….	



General VPM structure

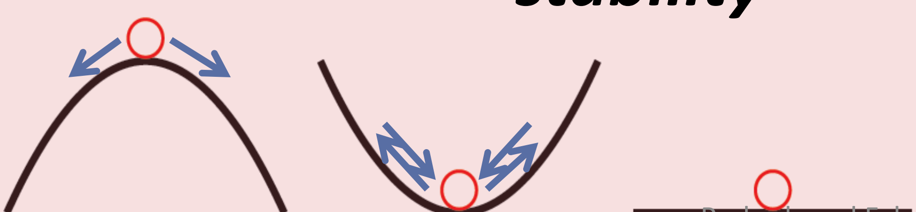
47

Model-based

$$\lambda * C_{VPM} < 1$$

Graph-based

Topology and stability

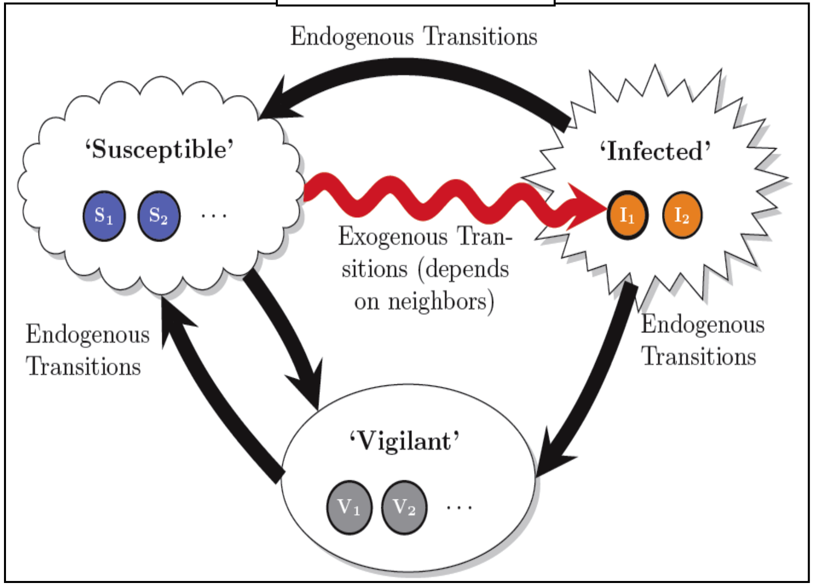


(A) Unstable (B) Stable (C) Neutral (at threshold)

Models and more models

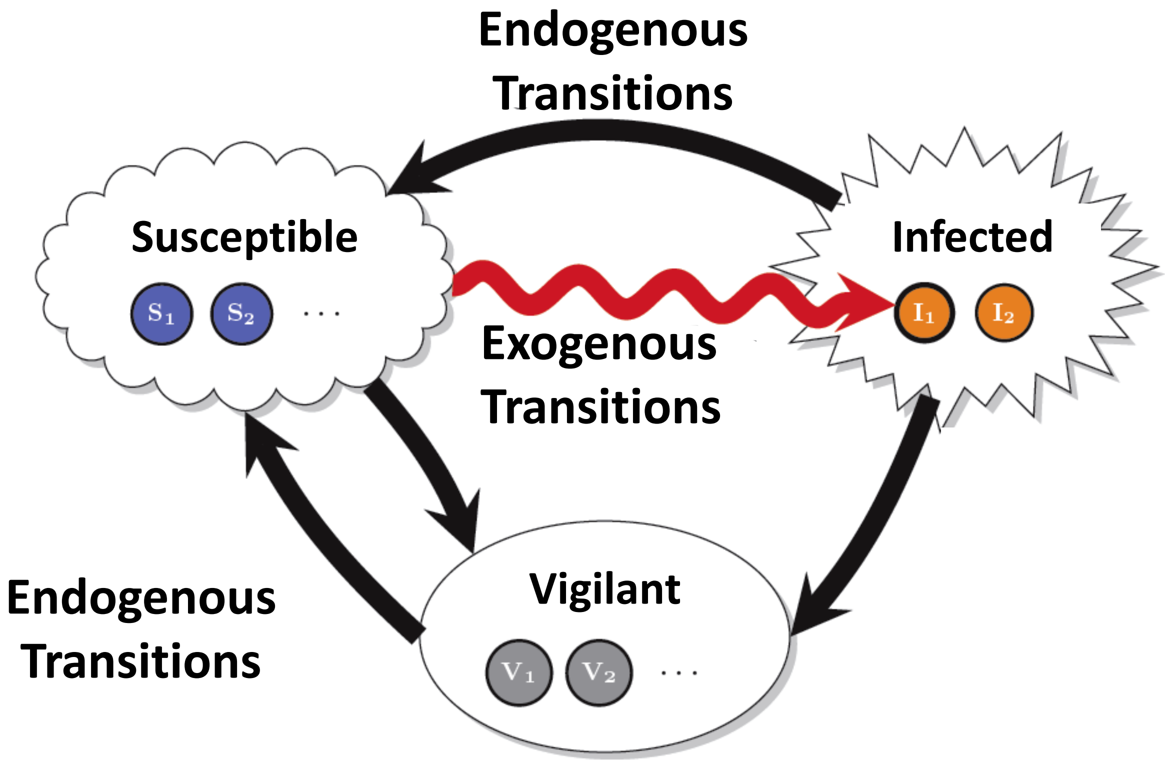
Model	Used for
SIR	Mumps
SIS	Flu
SIRS	Pertussis
SEIR	Chicken-pox
……..	
SICR	Tuberculosis
MSIR	Measles
SIV	Sensor Stability
$SI_1I_2V_1V_2$	H.I.V.
……….	

$S^*I^2V^*$

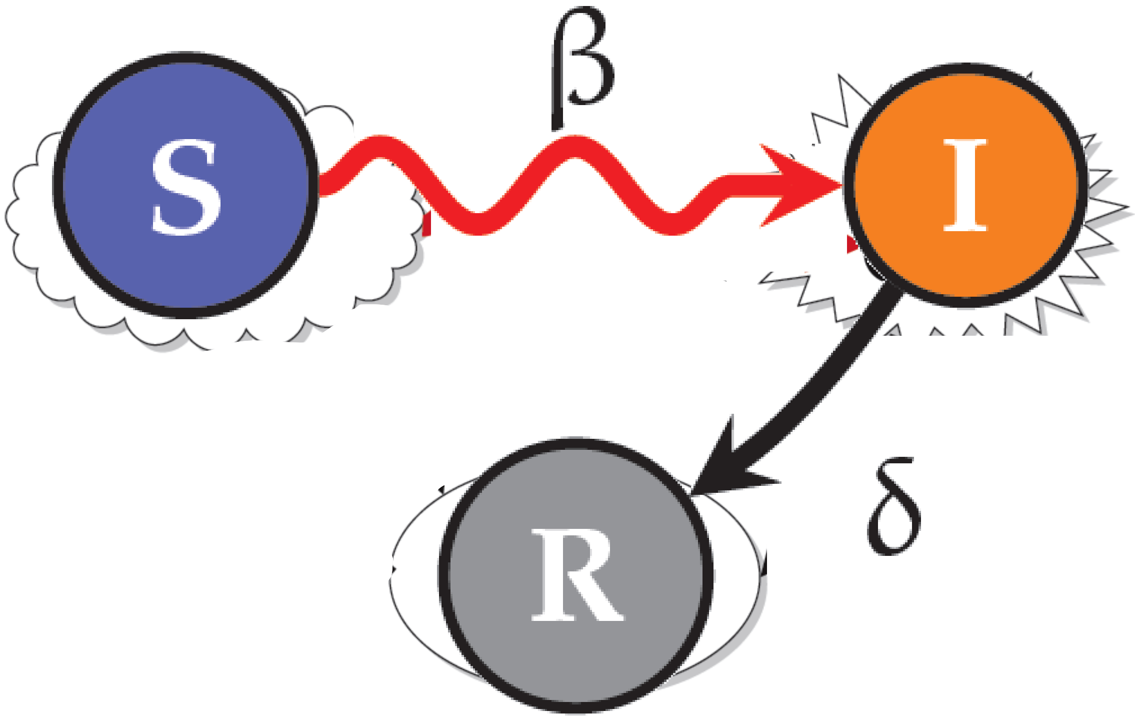


Ingredient 1: Our generalized model

$$S^*I^2V^* \;\; (S^*I^*V^*?)$$

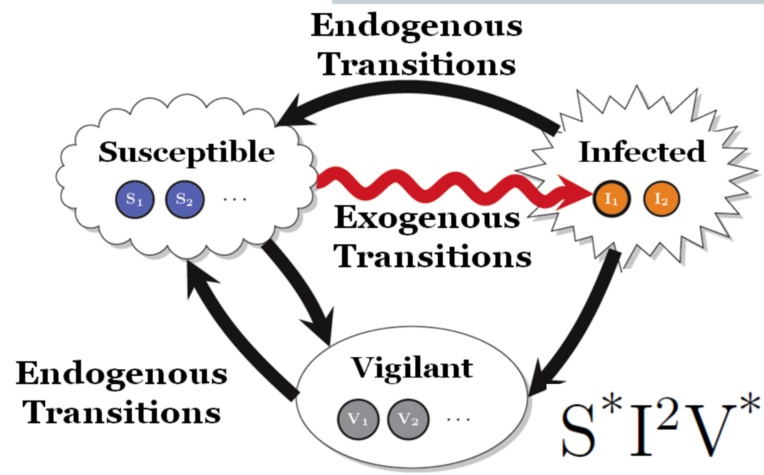


Special case: SIR

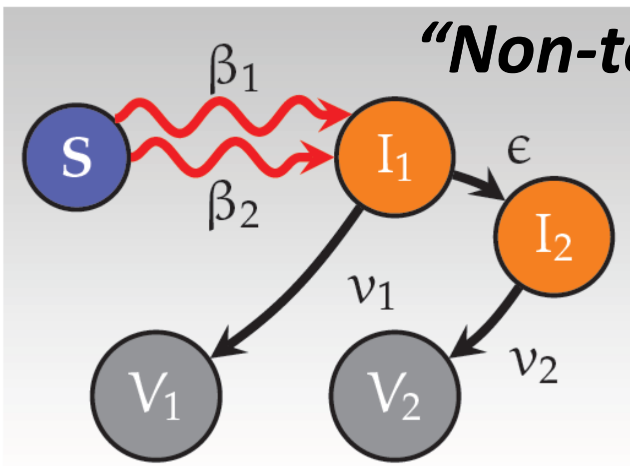


Special case: H.I.V.

$SI_1I_2V_1V_2$



$S^*I^2V^*$



"Non-terminal"

"Terminal"

Multiple Infectious, Vigilant states

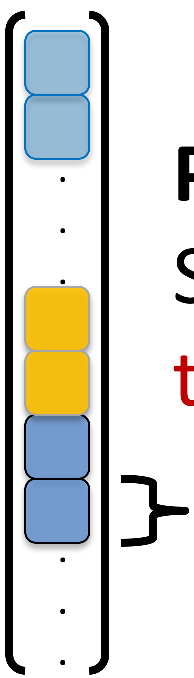
Details

Ingredient 2: NLDS+Stability

- View as a NLDS $\vec{P}_{t+1} = \mathcal{G}(\vec{P}_t)$
  - discrete time
  - non-linear dynamical system (NLDS)

$$\vec{P}_t$$

size
mN x *1*



Probability vector
Specifies the state of the system at time *t*

size N (number of nodes in the graph)

Details

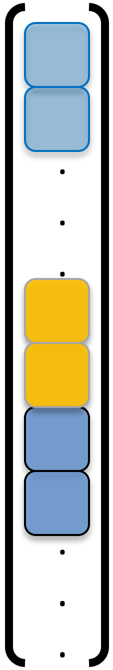
Ingredient 2: NLDS + Stability

- View as a NLDS $\vec{P}_{t+1} = \mathcal{G}(\vec{P}_t)$
  - discrete time
  - non-linear dynamical system (NLDS)

$$\vec{P}_t$$

size

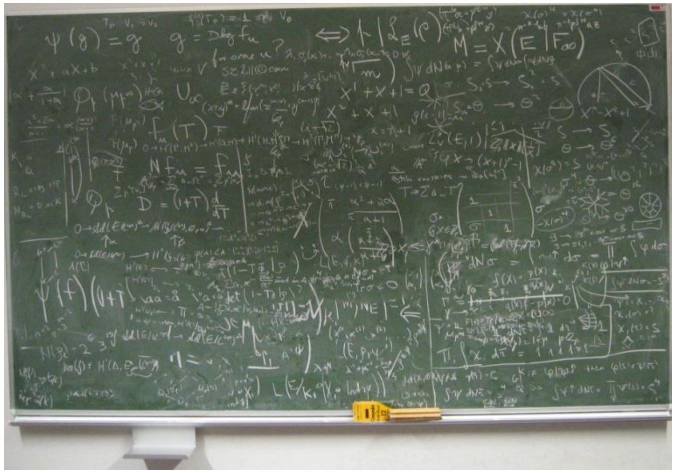
mN x *1*



$$\mathcal{G} : \mathbb{R}^{mN} \rightarrow \mathbb{R}^{mN}$$

Non-linear function

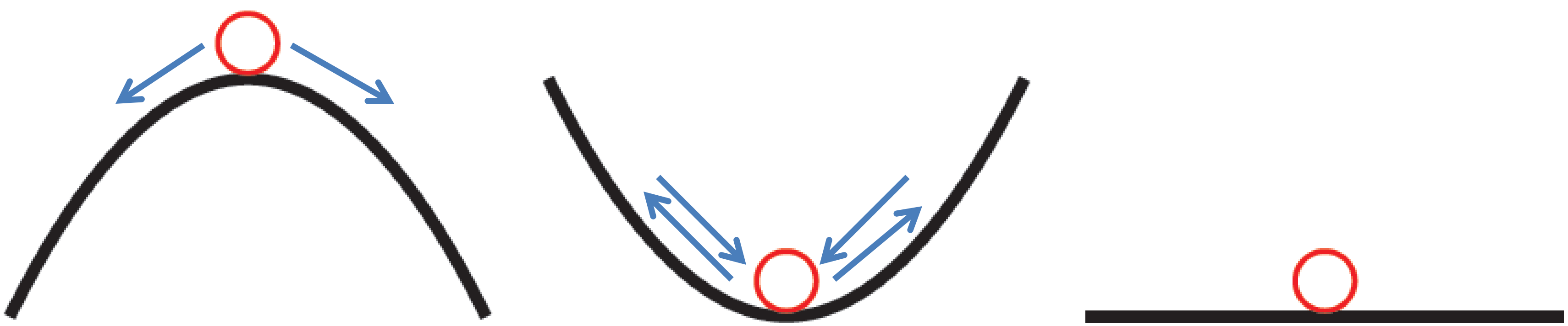
Explicitly gives the evolution of system



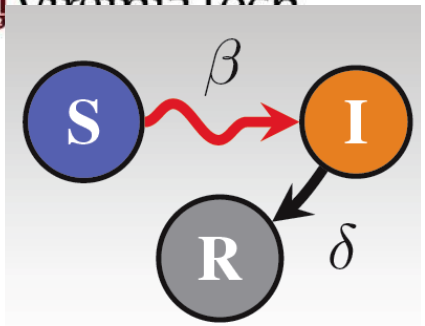
Ingredient 2: NLDS + Stability

- View as a NLDS $\vec{P}_{t+1} = \mathcal{G}(\vec{P}_t)$
  - discrete time
  - non-linear dynamical system (NLDS)

- Threshold → Stability of NLDS



(A) Unstable (B) Stable (C) Neutral (at threshold)



Details

Special case: SIR

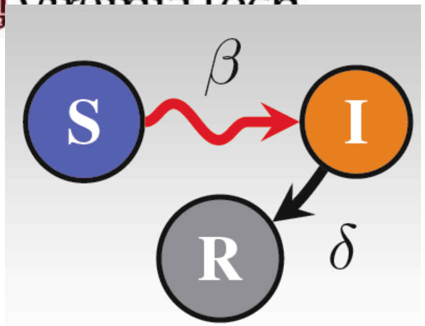
$$\mathcal{G} : \mathbb{R}^{3N} \rightarrow \mathbb{R}^{3N}$$

$$\vec{P}_{t+1} = \begin{bmatrix} S \\ I \\ R \end{bmatrix} = \boxed{\begin{aligned} P_{S,i,t+1} &= P_{S,i,t}\zeta_{i,t}(I) \\ P_{I,i,t+1} &= P_{S,i,t}(1-\zeta_{i,t}(I)) + (1-\delta)P_{I,i,t} \\ P_{R,i,t+1} &= \delta P_{I,i,t} + P_{R,i,t} \end{aligned}} \leftarrow \begin{bmatrix} S \\ I \\ R \end{bmatrix} \vec{P}_t$$

size

3N x *1*

$\zeta_{i,t}(I)$ = probability that node *i* is not attacked by any of its infectious neighbors



Details

Special case: SIR

$$\vec{P}_{t+1}$$

size

3N x *1*

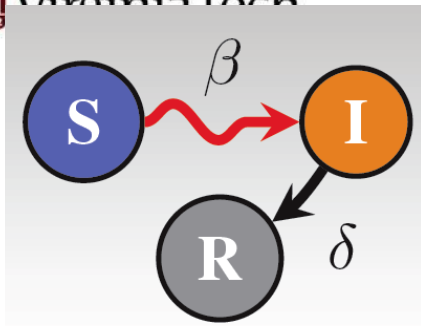
$$\begin{bmatrix} \text{S} \\ \text{I} \\ \text{R} \end{bmatrix} = $$

$$\mathcal{G} : \mathbb{R}^{3N} \rightarrow \mathbb{R}^{3N}$$

$$\boxed{\begin{aligned} P_{S,i,t+1} &= P_{S,i,t}\zeta_{i,t}(I) \\ P_{I,i,t+1} &= P_{S,i,t}(1-\zeta_{i,t}(I)) + (1-\delta)P_{I,i,t} \\ P_{R,i,t+1} &= \delta P_{I,i,t} + P_{R,i,t} \end{aligned}}$$

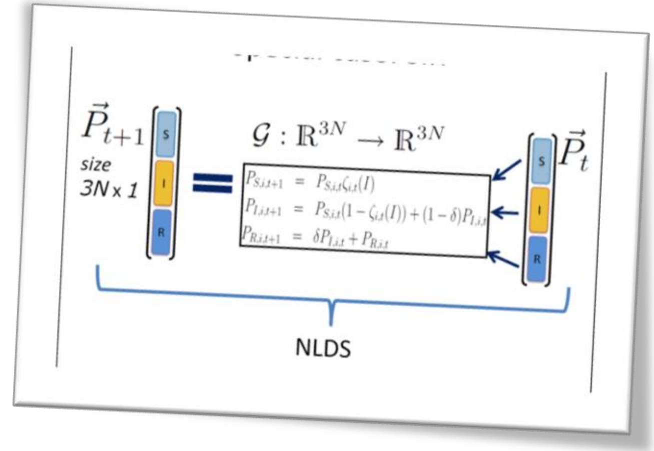
$$\begin{bmatrix} \text{S} \\ \text{I} \\ \text{R} \end{bmatrix} \vec{P}_t$$

$$\underbrace{\qquad\qquad\qquad\qquad\qquad\qquad}_{\text{NLDS}}$$



Details

Fixed Point

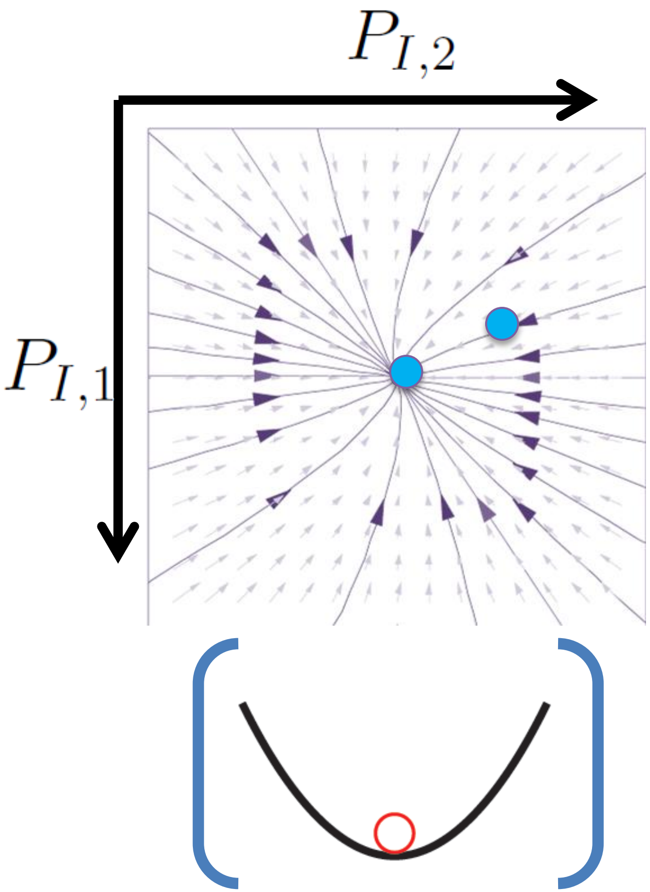


$$\vec{P}^* \begin{bmatrix} 1 \\ 1 \\ . \\ 0 \\ 0 \\ . \\ 0 \\ 0 \\ . \end{bmatrix}$$

State when no node is infected

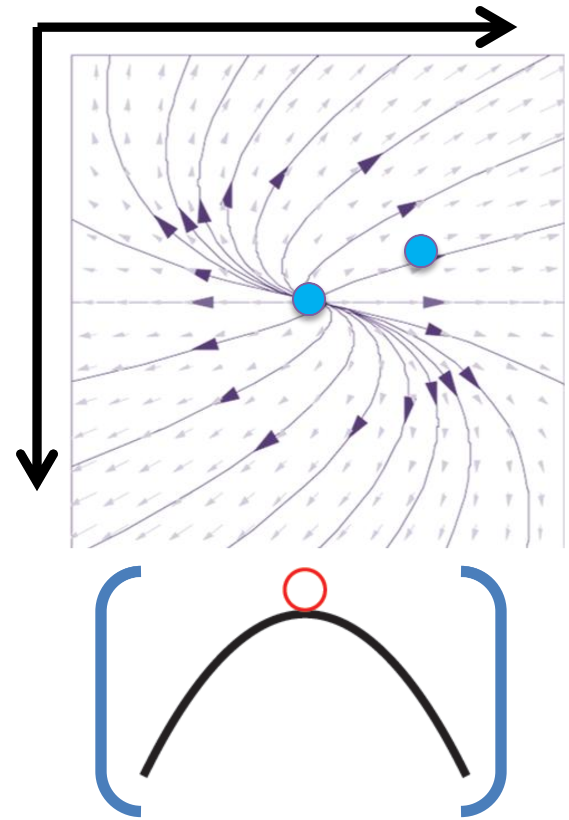
Q: Is it stable?

Stability for SIR



Stable

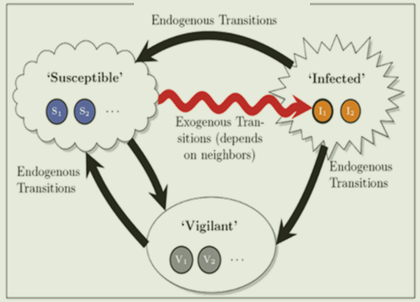
under threshold



Unstable

above threshold

Model	Used for
SIR	Mumps
SIS	Flu
SIRS	Pertussis
SEIR	Varicella
……..	
SICR	Tuberculosis
MSIR	Measles
SIV	Sensor Stability
$SI_1I_2V_1V_2$	H.I.V.
……….	



General VPM structure

47

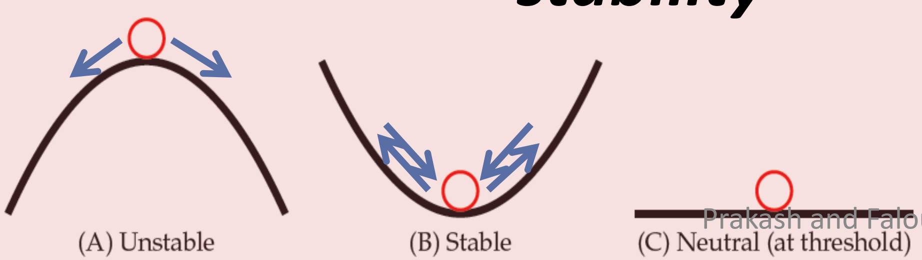
See paper for full proof

Model-based

$$\lambda * C_{VPM} < 1$$

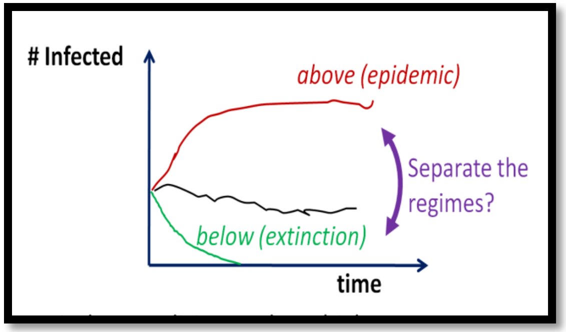
Graph-based

Topology and stability



Part 1: Theory

- Q1: What is the epidemic threshold?
  - Background
  - Result and Intuition (Static Graphs)
  - Proof Ideas (Static Graphs)
  - **Bonus: Dynamic Graphs**

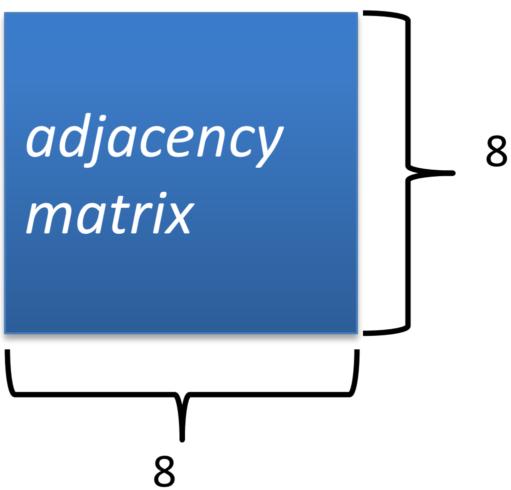


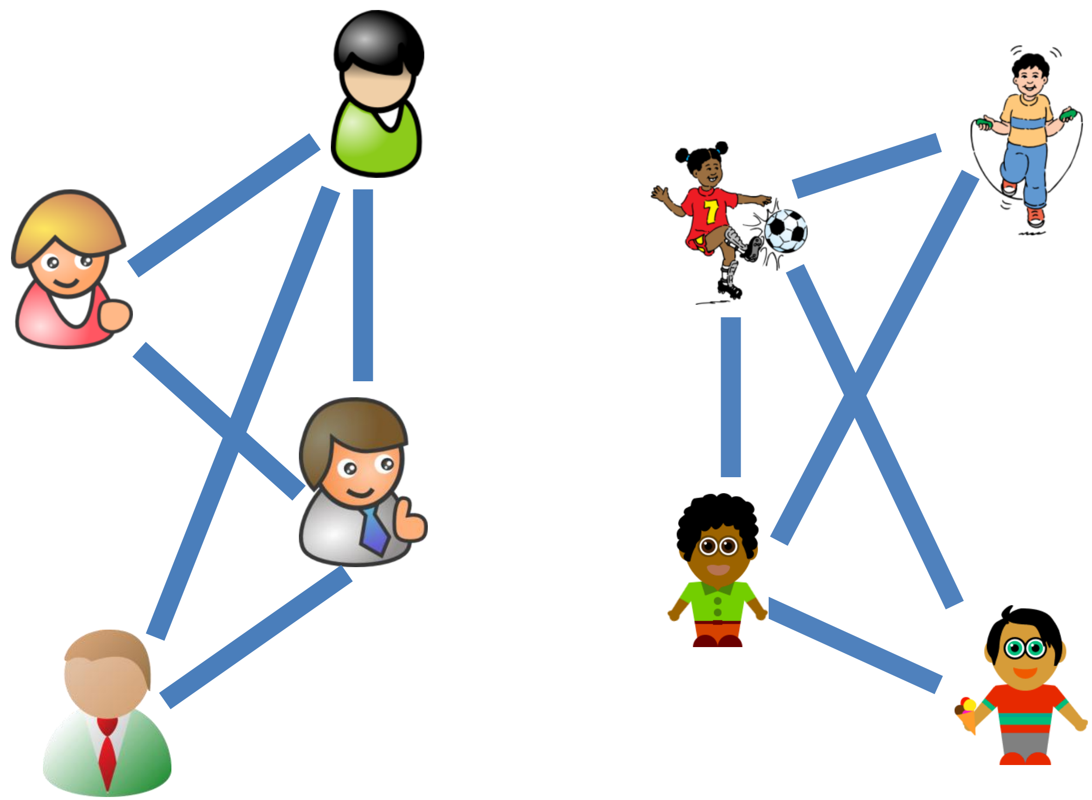
- Q2: How do viruses compete?

Dynamic Graphs: Epidemic?

Alternating behaviors

DAY
(e.g., work)

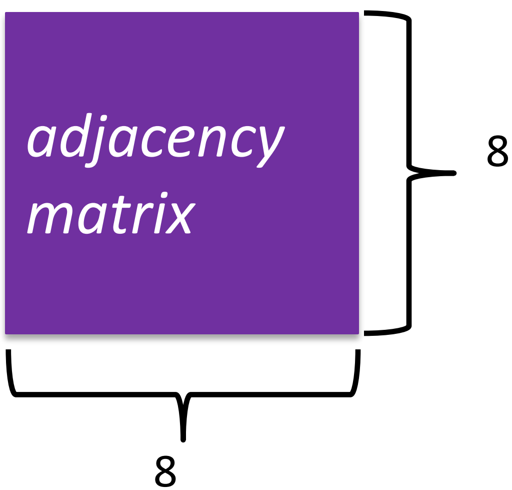


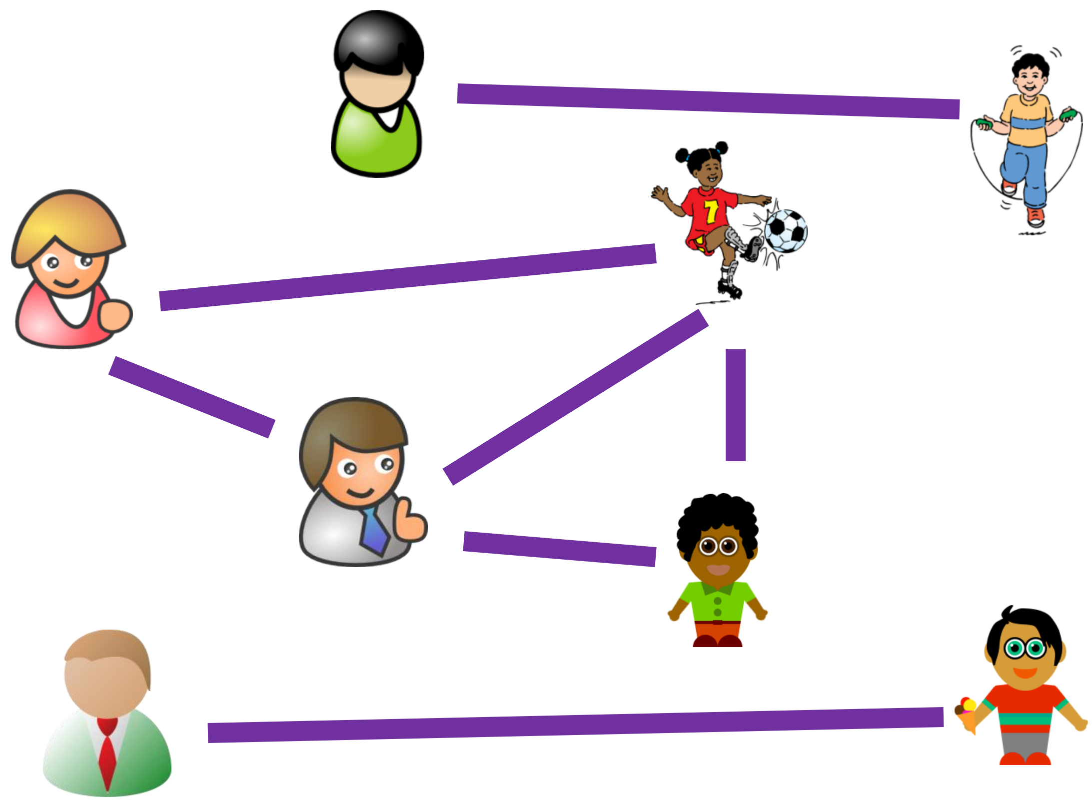


Dynamic Graphs: Epidemic?

NIGHT
(e.g., home)

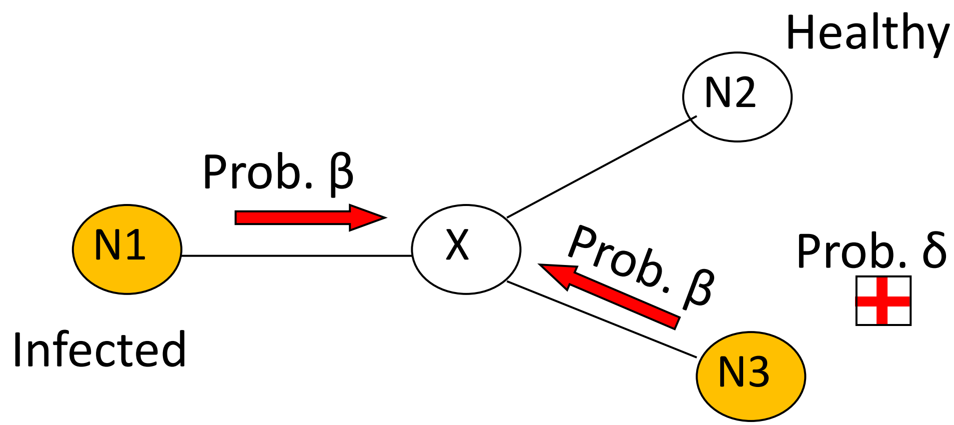
Alternating behaviors





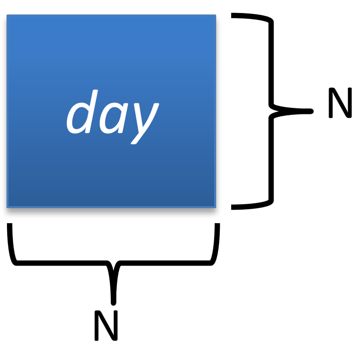
Model Description

- SIS model
  - recovery rate δ
  - infection rate β

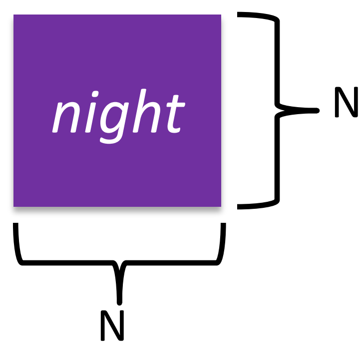


- Set of *T* **arbitrary** graphs $\{\mathbf{A}_1, \mathbf{A}_2 \ \ldots, \mathbf{A}_T\}$

$\mathbf{A}_1$



$\mathbf{A}_2$



, weekend…..

Our result: Dynamic Graphs Threshold

- **Informally, *NO* epidemic if**

$$\textit{eig}\ (\mathbf{S}) = \lambda_{\mathbf{S}} < 1$$

Single number!
Largest eigenvalue of
The *system matrix* S

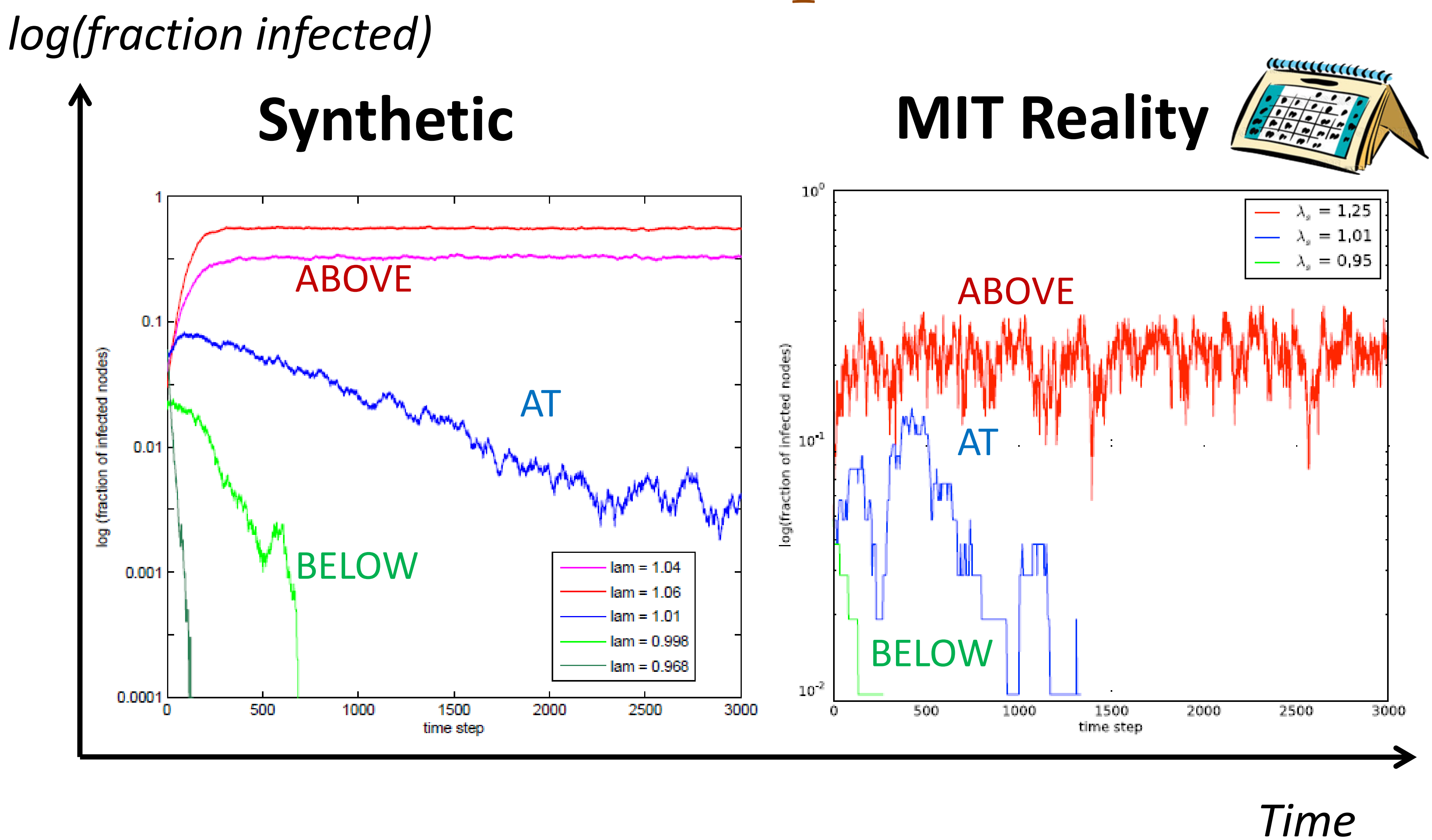
Details

$$\mathbf{S} = \prod_i \mathbf{S}_i$$

$$\mathbf{S}_i = (1-\delta)\mathbf{I} + \beta \mathbf{A}_i$$

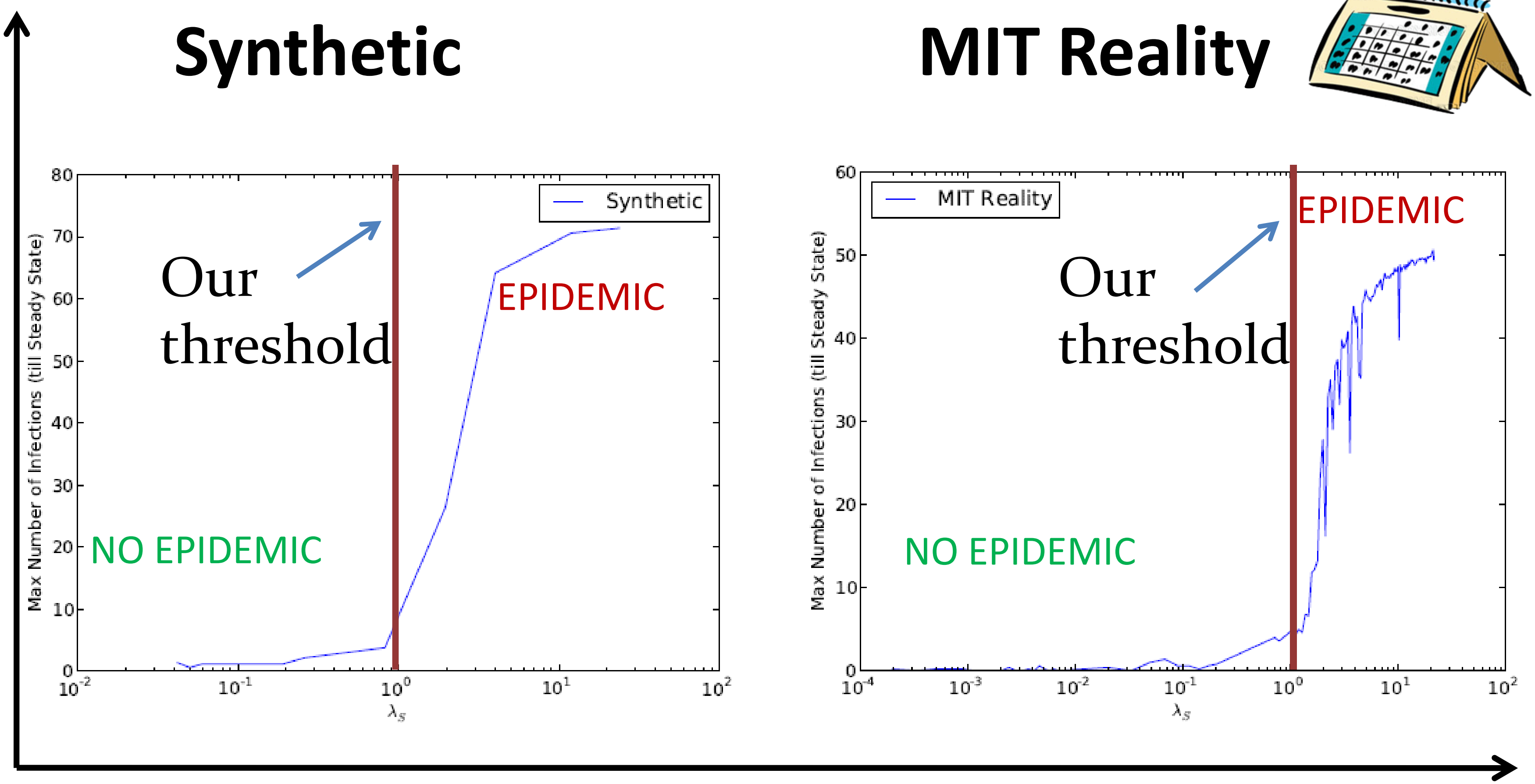
In Prakash+, ECML-PKDD 2010

Infection-profile



Footprint (# infected @ "steady state")

"Take-off" plots



$\lambda \prod_i \mathbf{s}_i$ *(log scale)*

Part 1: Theory

- Q1: What is the epidemic threshold?
- **Q2: What happens when viruses compete?**
  - **Mutually-exclusive viruses**
  - Interacting viruses

Competing Contagions

iPhone v **Android**

Blu-ray v **HD-DVD**

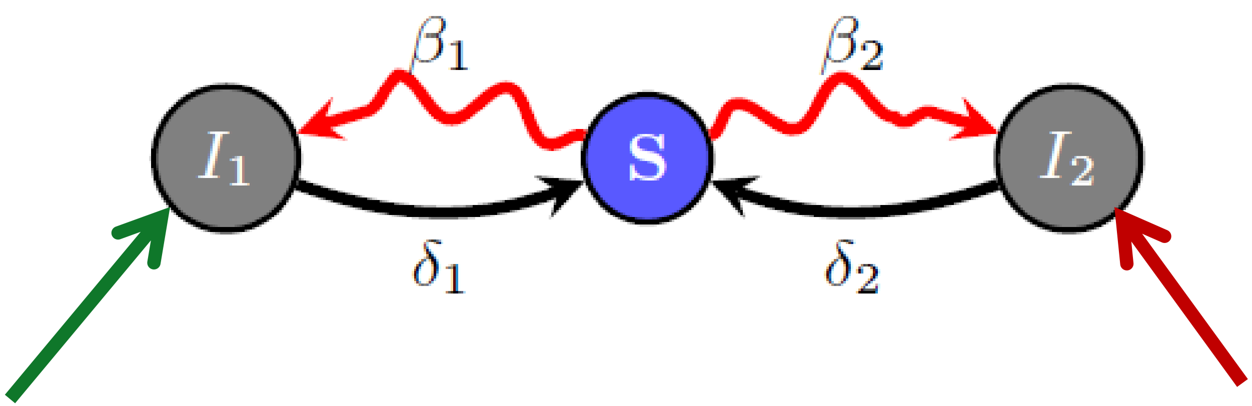
Attack v **Retreat**

Biological common flu/avian flu, pneumococcal inf etc

Details

A simple model

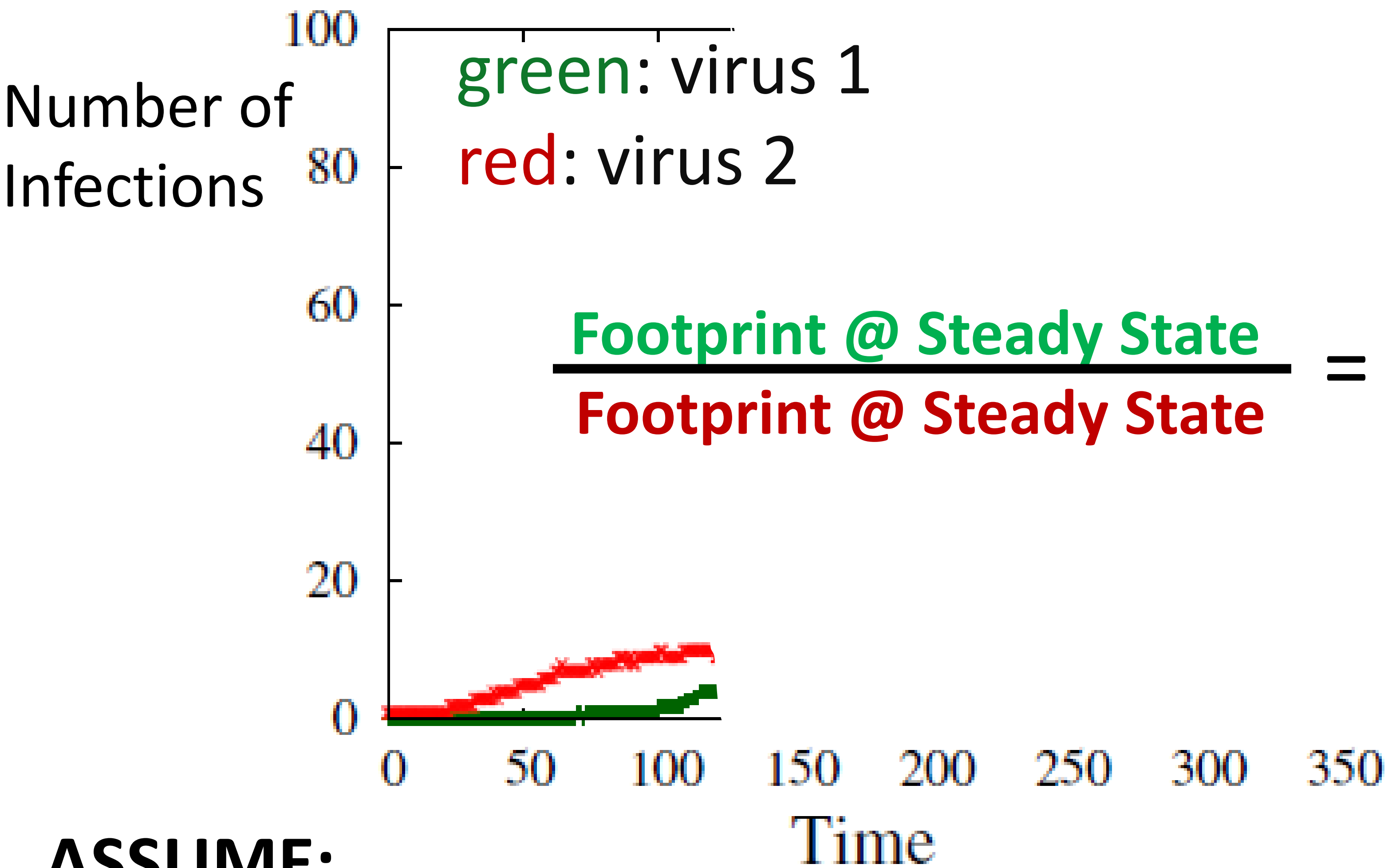
- Modified flu-like
- Mutual Immunity (“pick one of the two”)
- Susceptible-Infected1-Infected2-Susceptible



Virus 1

Virus 2

Question: What happens in the end?

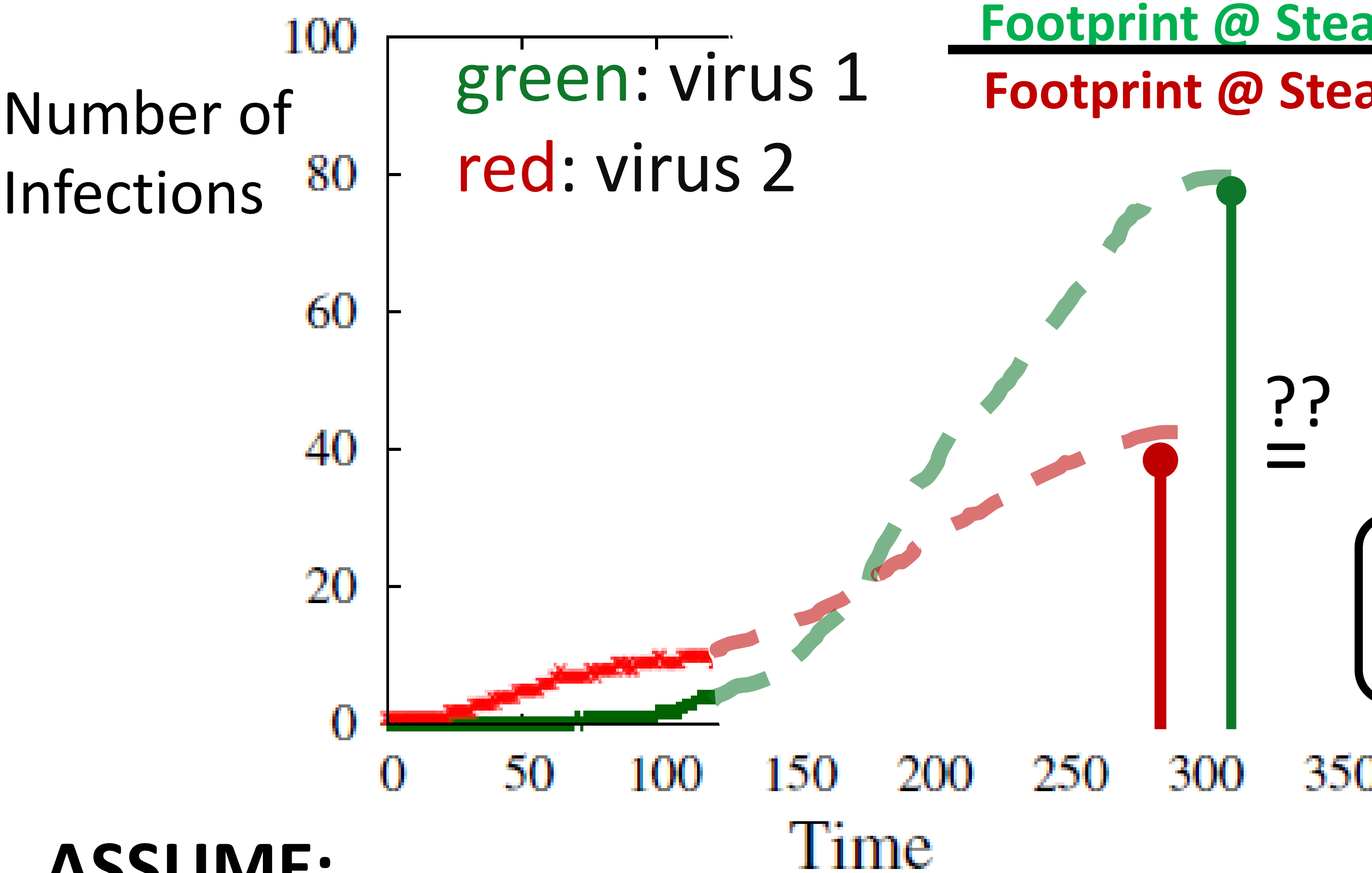


$$\frac{\text{Footprint @ Steady State}}{\text{Footprint @ Steady State}} = \ ?$$

ASSUME:

Virus 1 is stronger than Virus 2

Question: What happens in the end?

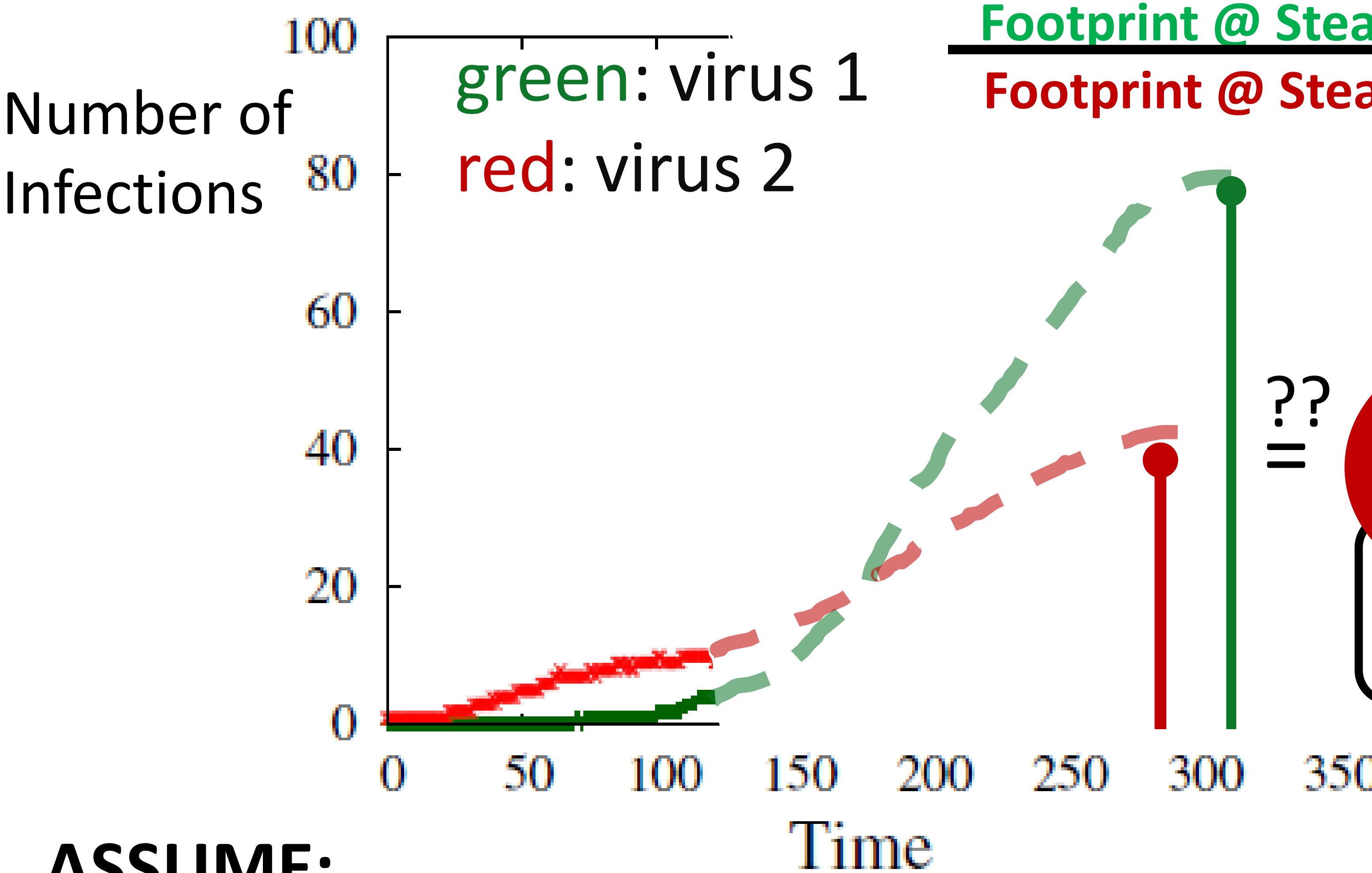


$$\frac{\text{Footprint @ Steady State}}{\text{Footprint @ Steady State}} \overset{??}{=} \frac{\textit{Strength}}{\textit{Strength}}$$

$$\left(\frac{\textit{Strength}}{\textit{Strength}}\right)^2$$

ASSUME:
Virus 1 is stronger than Virus 2

Question: What happens in the end?

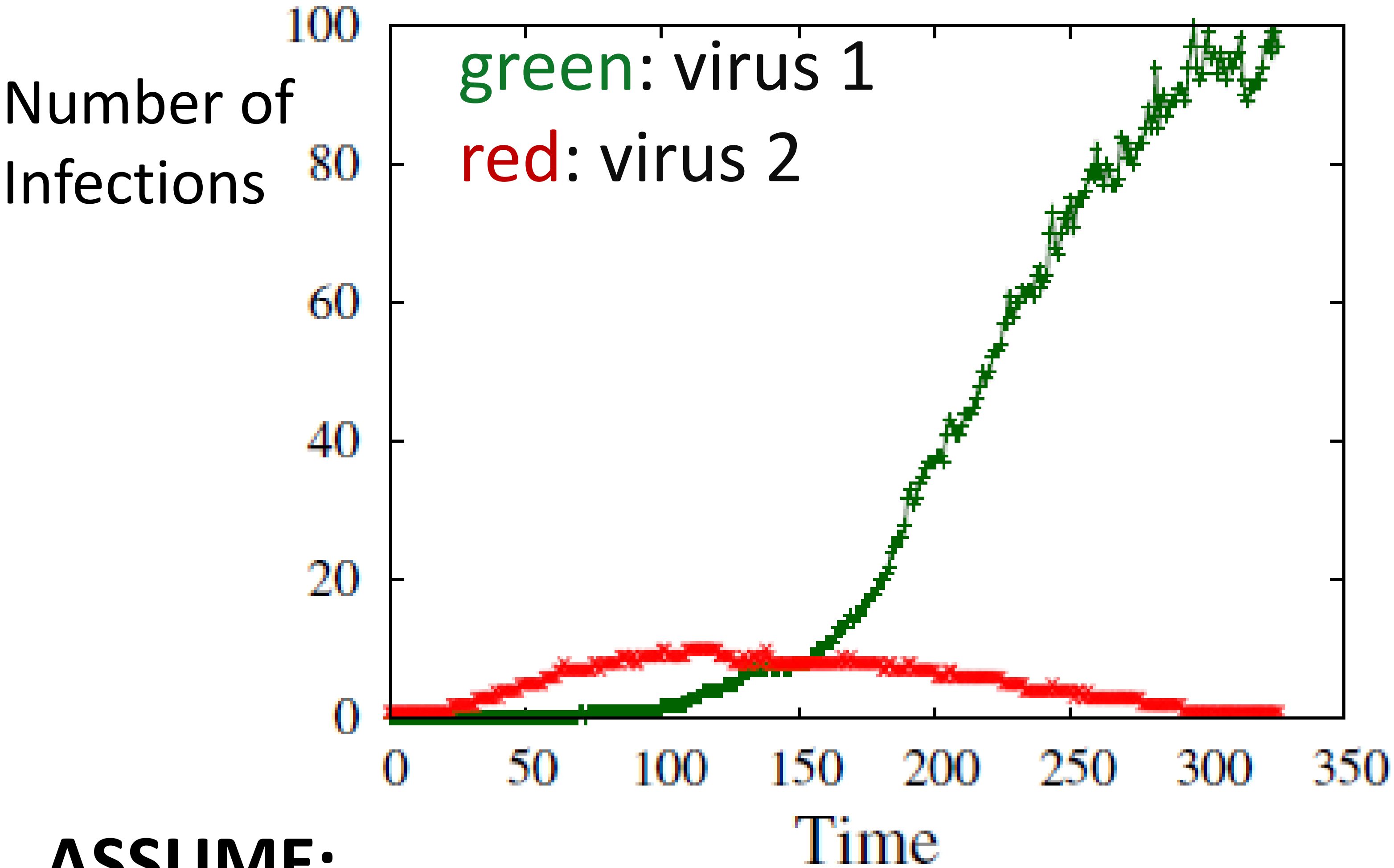


$$\frac{\text{Footprint @ Steady State}}{\text{Footprint @ Steady State}} \overset{??}{=} \cancel{\frac{\text{Strength}}{\text{Strength}}} \quad \cancel{\left(\frac{\text{Strength}}{\text{Strength}}\right)^2}$$

ASSUME:

Virus 1 is stronger than Virus 2

Answer: Winner-Takes-All



ASSUME:

Virus 1 is stronger than Virus 2

Our Result: Winner-Takes-All

Given our model, and ***any graph***, the weaker virus always **dies-out completely**

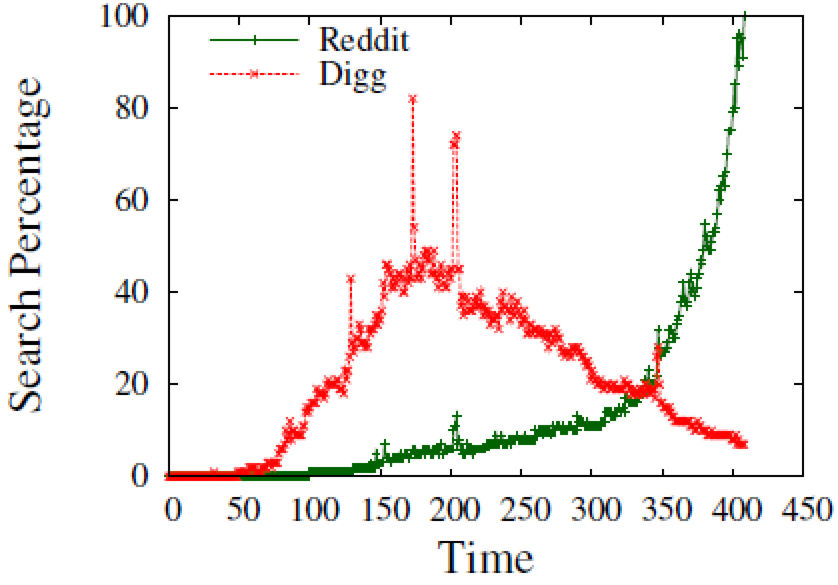
Details

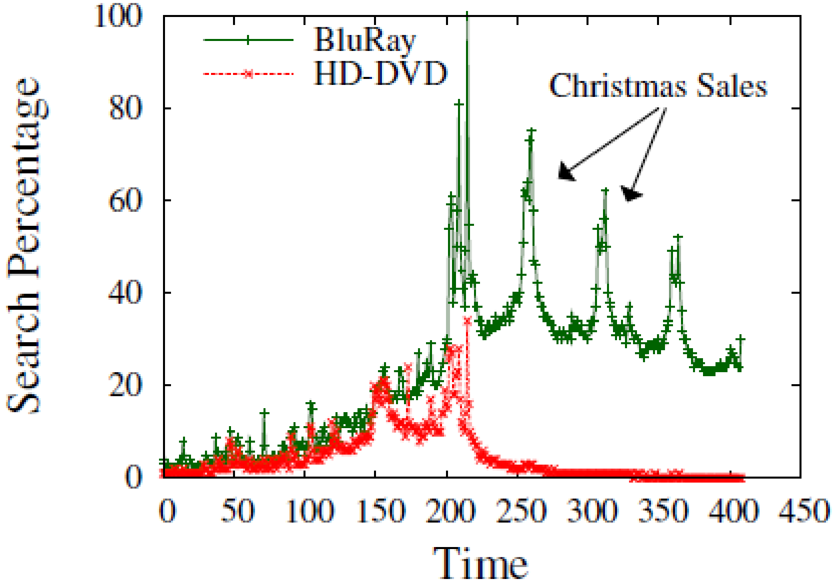
1. The stronger survives only if it is above threshold
2. Virus 1 is stronger than Virus 2, if:
   strength(Virus 1) > strength(Virus 2)
3. Strength(Virus) = $\lambda\ \beta\ /\ \delta$ → same as before!

In Prakash+ WWW 2012

Real Examples

[Google Search Trends data]





Reddit v **Digg**

Blu-Ray v **HD-DVD**

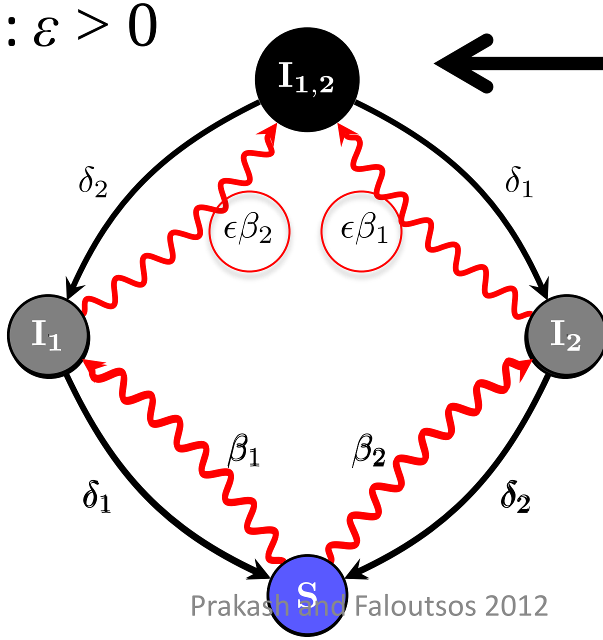
Part 1: Theory

- Q1: What is the epidemic threshold?
- **Q2: What happens when viruses compete?**
  - Mutually-exclusive viruses
  - **Interacting viruses**

A simple model: $SI_{1|2}S$

- Modified flu-like (SIS)
- Susceptible-Infected$_{1 \text{ or } 2}$-Susceptible
- Interaction Factor ε
  - Full Mutual Immunity: $\varepsilon = 0$
  - Partial Mutual Immunity (competition): $\varepsilon < 0$
  - Cooperation: $\varepsilon > 0$

&



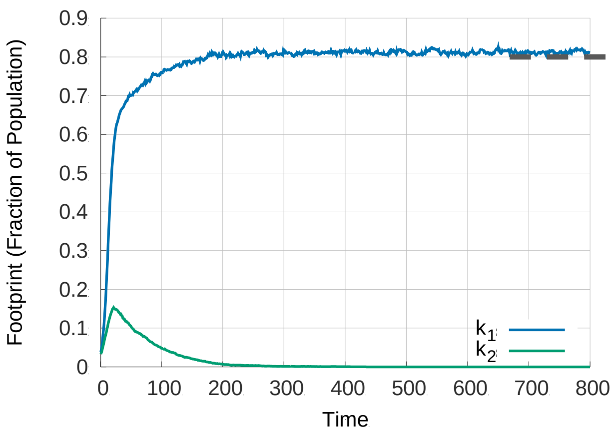
Virus 1

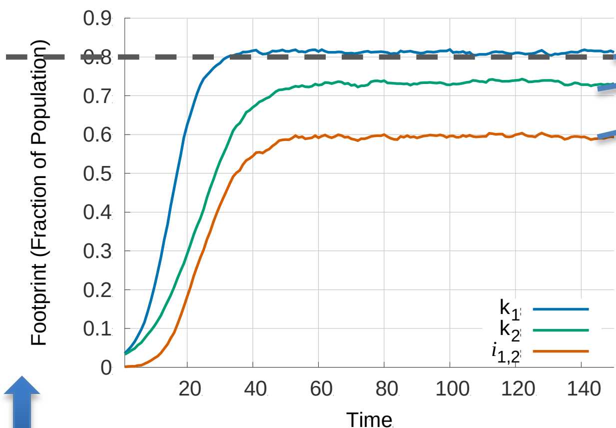
Virus 2

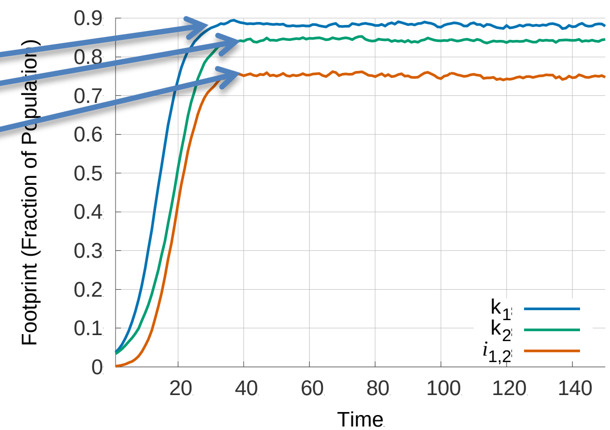
Prakash and Faloutsos 2012

Question: What happens in the end?

$\varepsilon = 0$	$\varepsilon = 1$	$\varepsilon = 2$
Winner takes all	Co-exist independently	Viruses cooperate





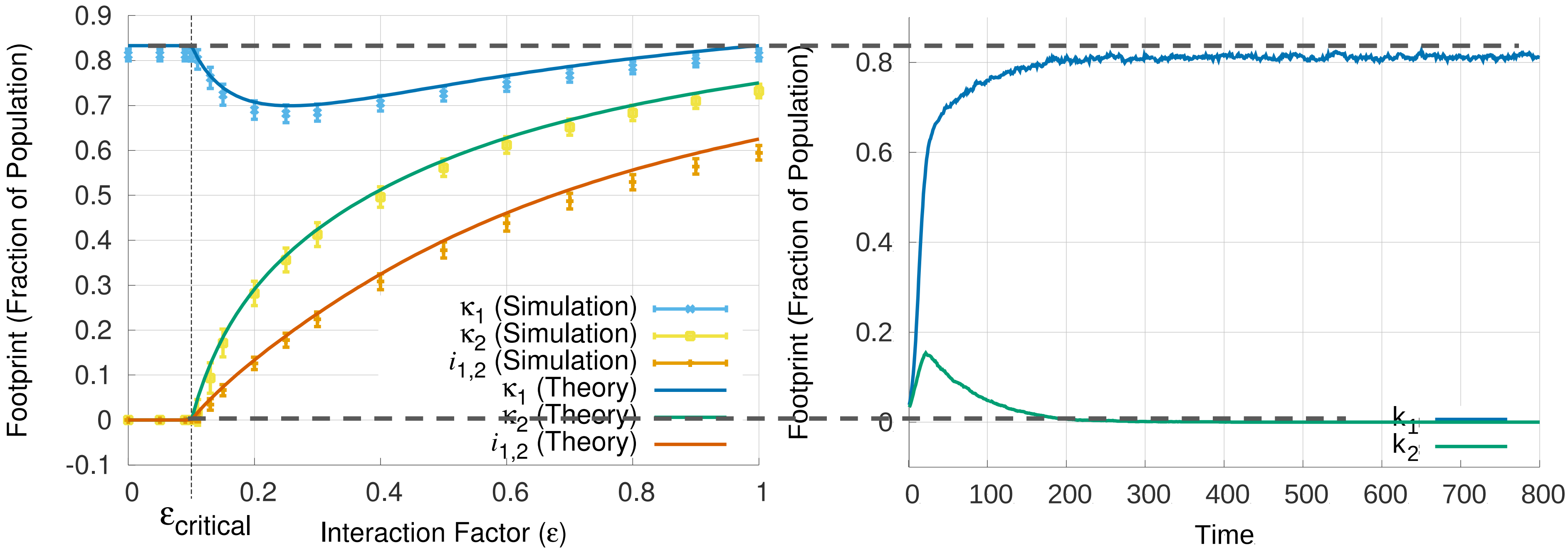


What about for $0 < \varepsilon < 1$?
Is there a point at which both viruses can *co-exist*?

ASSUME:
Virus 1 is stronger than Virus 2

Answer: Yes!

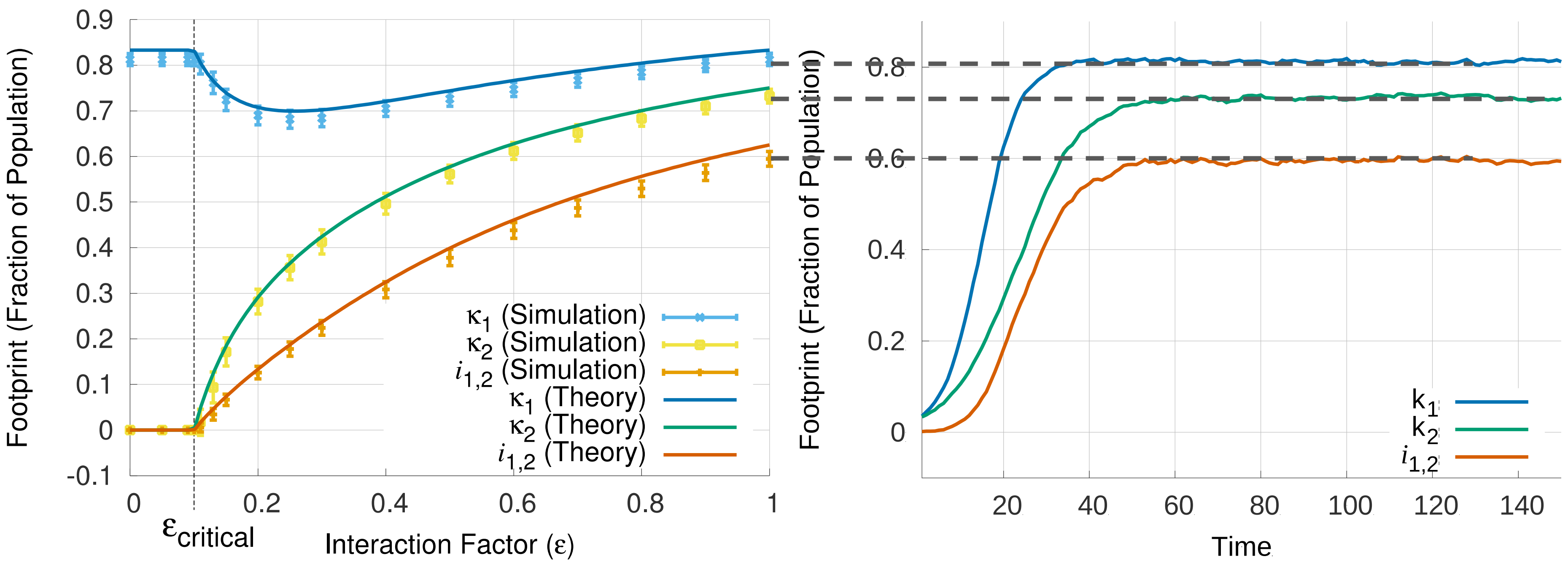
There is a phase transition



ASSUME:

Virus 1 is stronger than Virus 2

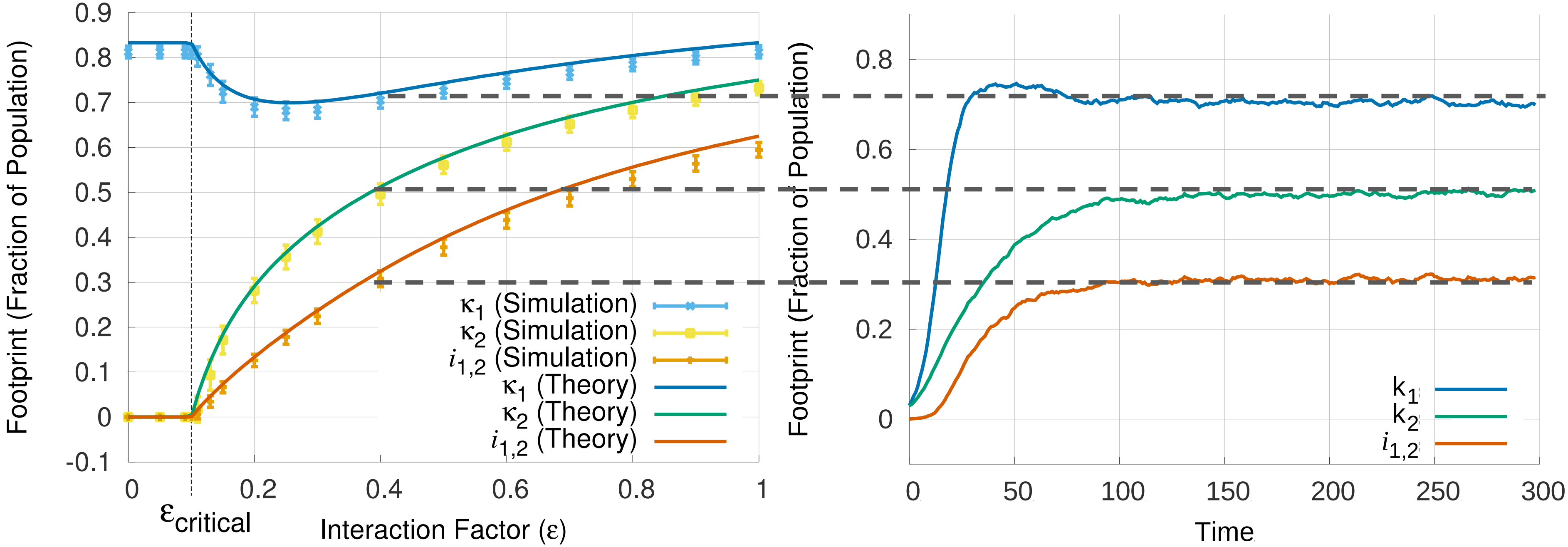
Answer: Yes!
There is a phase transition



ASSUME:

Virus 1 is stronger than Virus 2

Answer: Yes!
There is a phase transition



ASSUME:

Virus 1 is stronger than Virus 2

Our Result: Viruses can Co-exist

Given our model and a fully connected graph, there exists an $\varepsilon_{critical}$ such that for $\varepsilon \geq \varepsilon_{critical}$, there is a fixed point where both viruses survive.

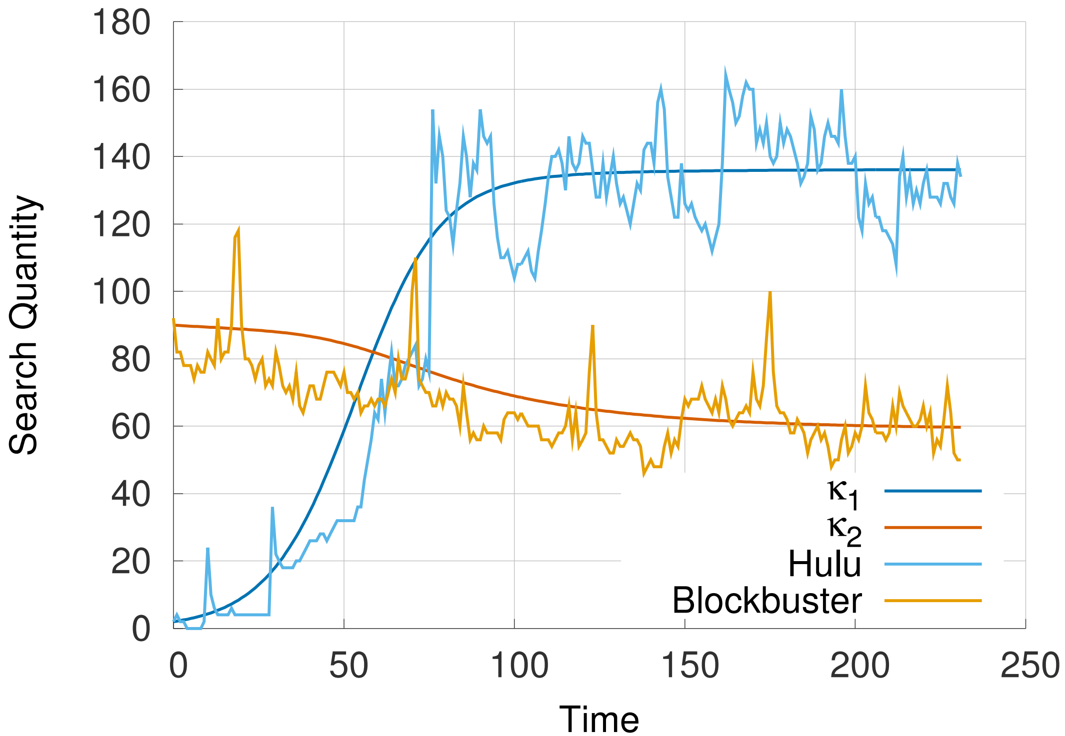
Details

1. The stronger survives only if it is above threshold
2. Virus 1 is stronger than Virus 2, if:
   strength(Virus 1) > strength(Virus 2)
3. Strength(Virus) $\sigma = N \beta / \delta$

In Beutel+ KDD 2012

Real Examples

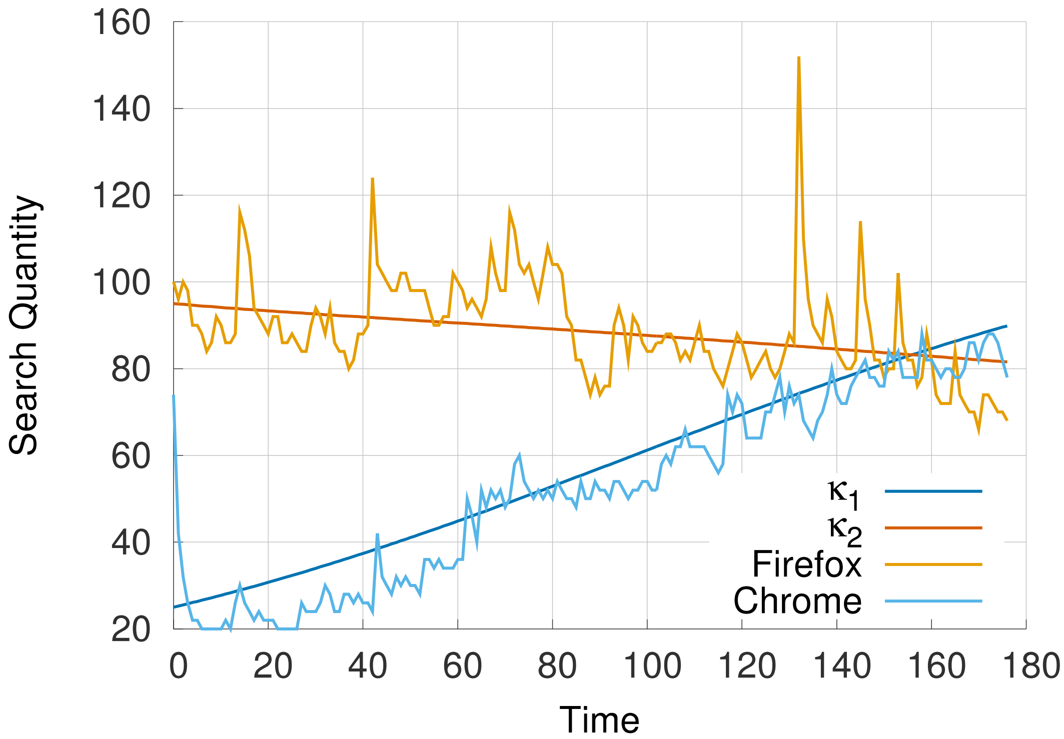
[Google Search Trends data]



Hulu v Blockbuster

Real Examples

[Google Search Trends data]



Chrome v **Firefox**

Outline

- Motivation
- Part 1: Understanding Epidemics (Theory)
- **Part 2: Policy and Action (Algorithms)**
- Part 3: Learning Models (Empirical Studies)
- Conclusion

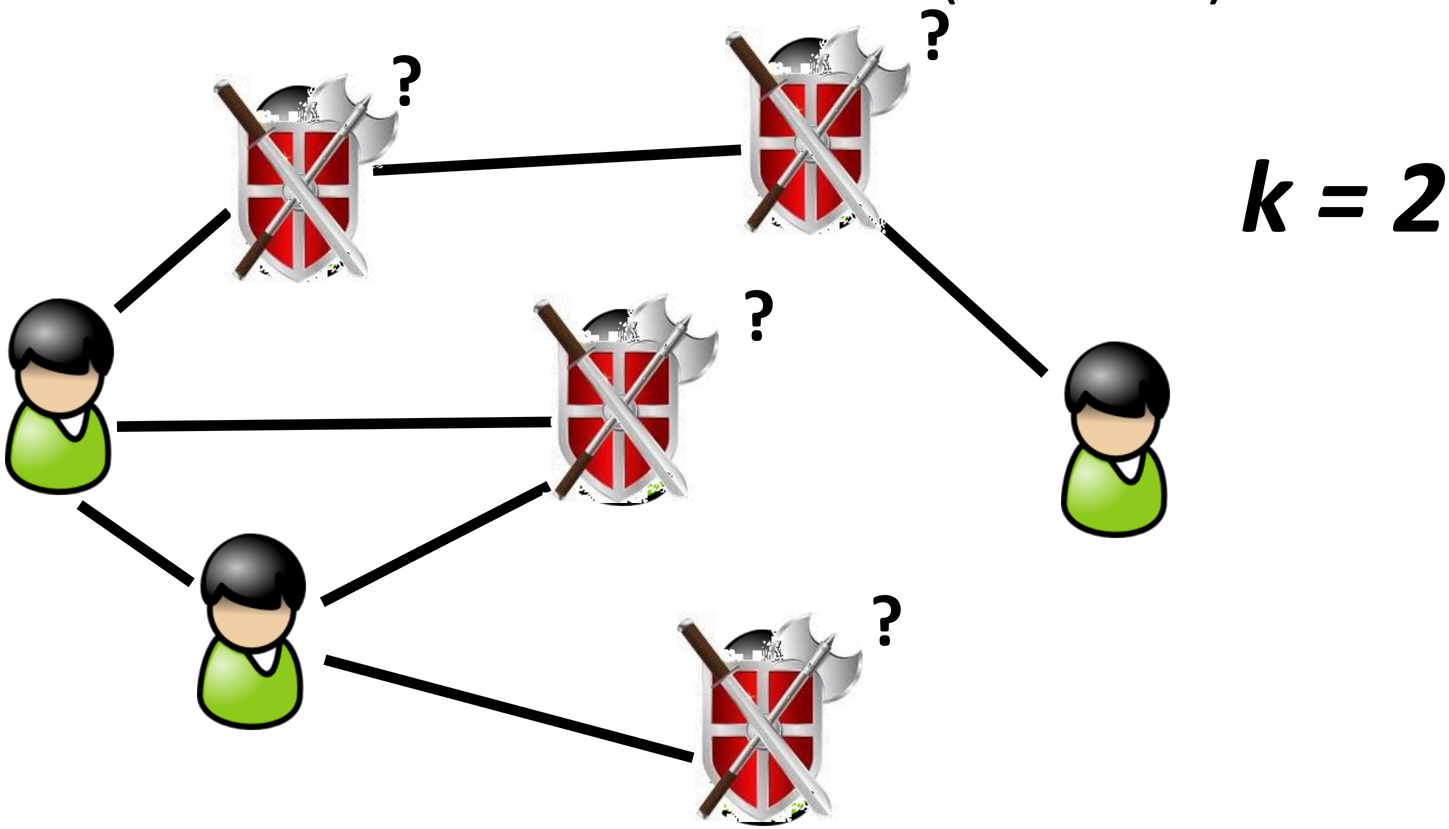
Part 2: Algorithms

- **Q3: Whom to immunize?**
- Q4: How to detect outbreaks?
- Q5: Who are the culprits?

Full Static Immunization

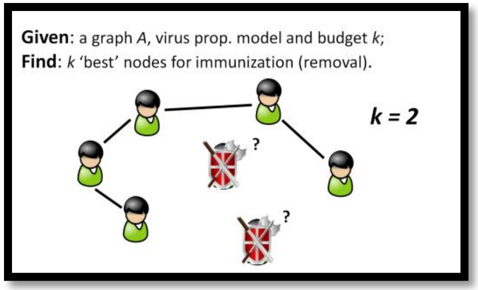
Given: a graph *A*, virus prop. model and budget *k*;

Find: *k* 'best' nodes for immunization (removal).



Part 2: Algorithms

- **Q3: Whom to immunize?**
  - **Full Immunization (Static Graphs)**
  - Full Immunization (Dynamic Graphs)
  - Fractional Immunization
- Q4: How to detect outbreaks?
- Q5: Who are the culprits?



Challenges

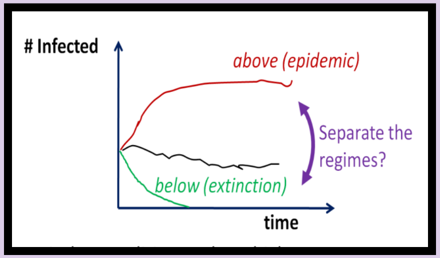
- Given a graph A, budget k,

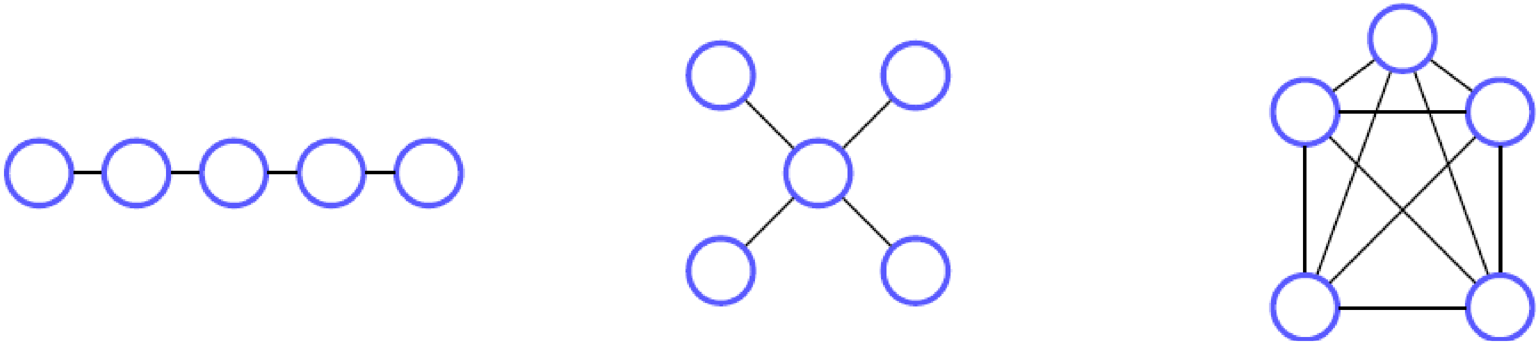
Q1 (Metric) How to measure the 'shield-value' for a set of nodes (S)?

Q2 (Algorithm) How to find a set of k nodes with highest 'shield-value'?

Proposed vulnerability measure λ

λ is the epidemic threshold





"Safe" *"Vulnerable"* *"Deadly"*

(a)Chain($\lambda = 1.73$) (b)Star($\lambda = 2$) (c)Clique($\lambda = 4$)

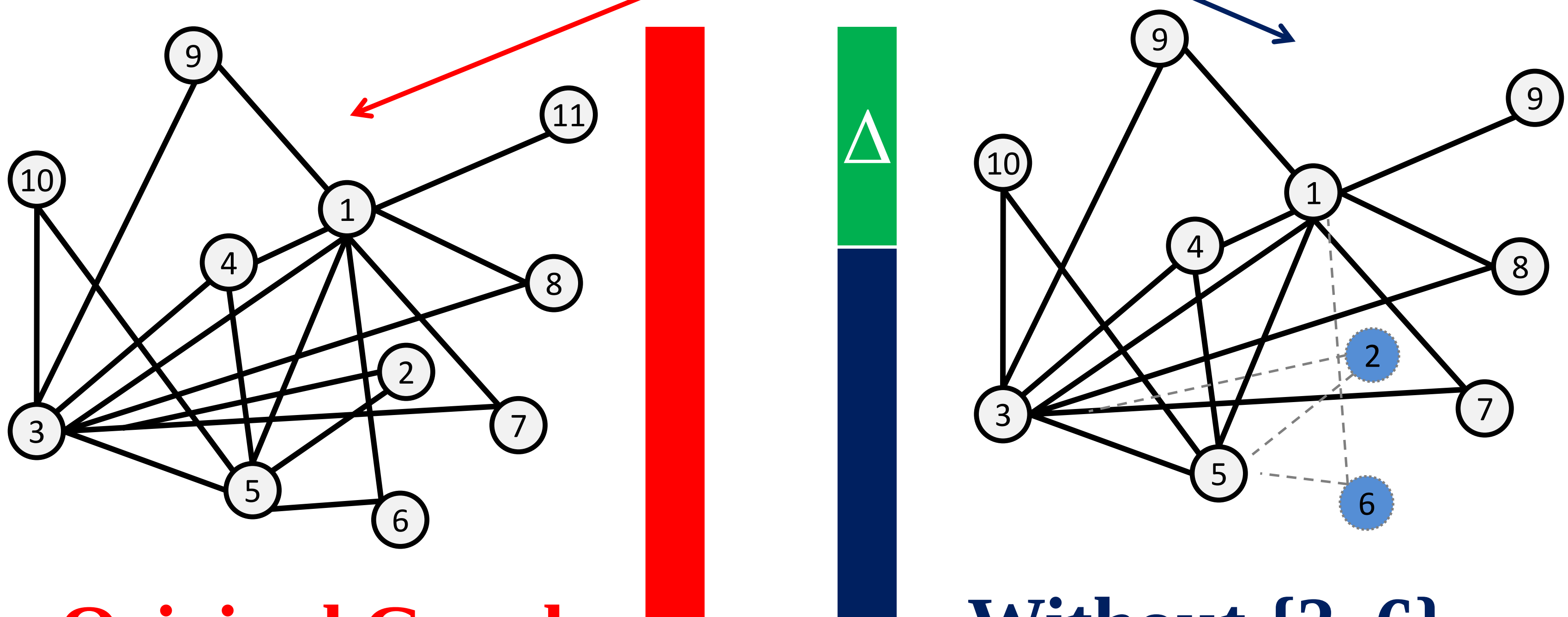
Increasing λ

Increasing vulnerability

A1: "Eigen-Drop": an ideal shield value

Eigen-Drop(S)

$$\Delta \lambda = \lambda - \lambda_s$$

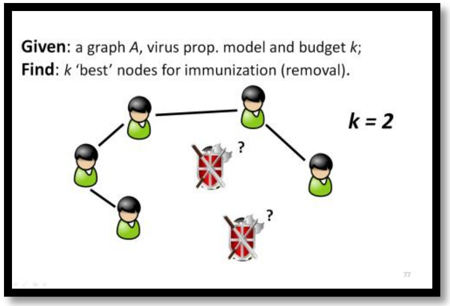


Original Graph **Without {2, 6}**

(Q2) - Direct Algorithm too expensive!

- Immunize k nodes which maximize $\Delta\lambda$

$$S = \text{argmax}\ \Delta\lambda$$



- Combinatorial!
- Complexity: $O(\binom{n}{k} \cdot m)$
  - Example:
    - 1,000 nodes, with 10,000 edges
    - It takes 0.01 seconds to compute λ
    - It takes **2,615 years** to find 5-best nodes!

A2: Our Solution

- Part 1: Shield Value
  - Carefully approximate Eigen-drop ($\Delta\,\lambda$)
  - Matrix perturbation theory
- Part 2: Algorithm
  - Greedily pick best node at each step
  - Near-optimal due to submodularity
- *NetShield* (linear complexity)
  - $O(nk^2+m)$ n = # nodes; m = # edges

Details

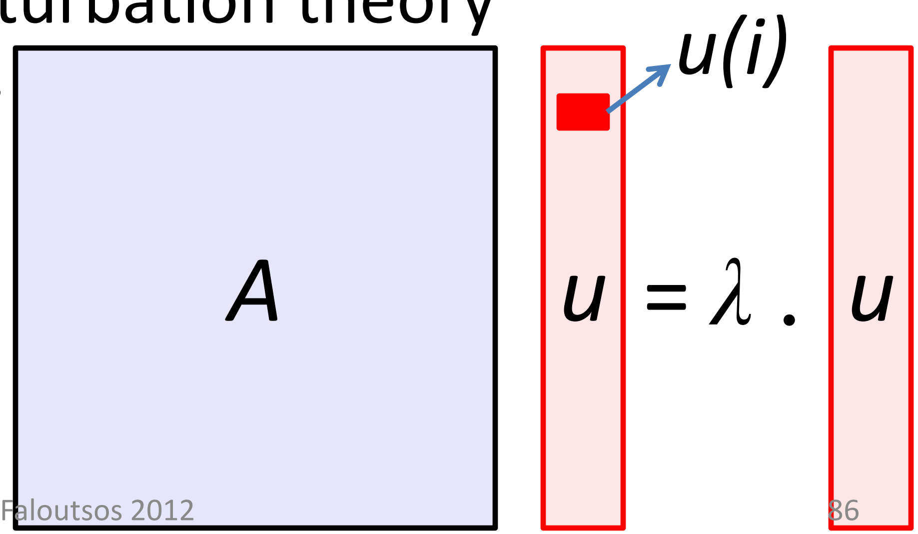
Our Solution: Part 1

- Approximate Eigen-drop ($\Delta\lambda$)

- $$\Delta\lambda \approx \widehat{SV}(S) = \sum_{i \in \mathcal{S}} 2\lambda \mathbf{u}(i)^2 - \sum_{i,j \in \mathcal{S}} \mathbf{A}(i,j)\mathbf{u}(i)\mathbf{u}(j)$$

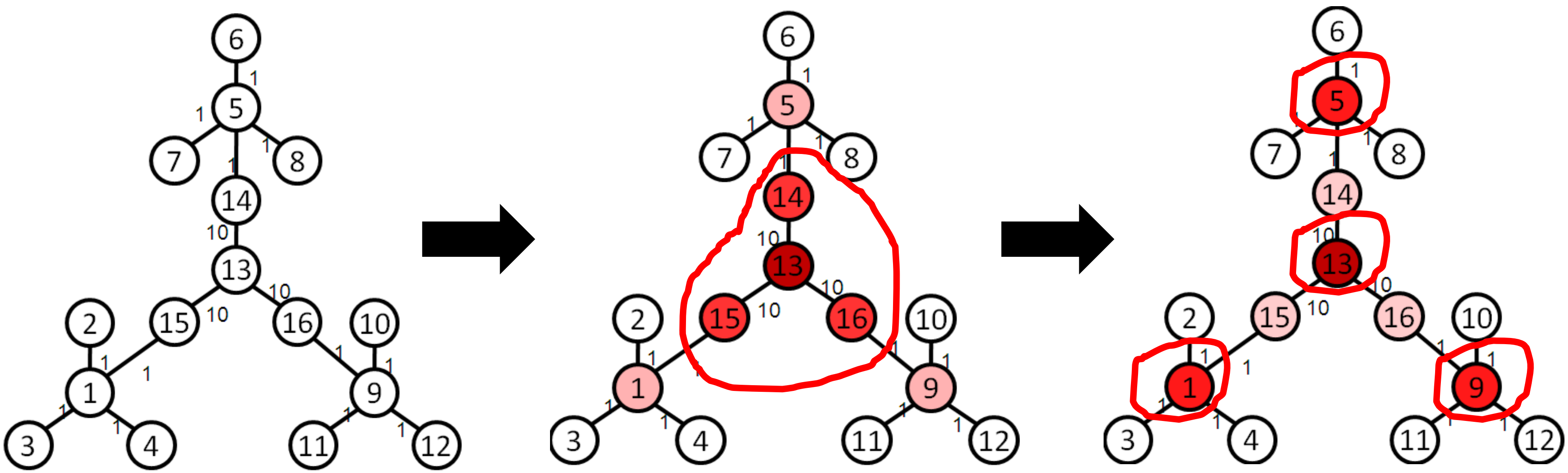
  - Result using Matrix perturbation theory
  - *u(i) == 'eigenscore'*

    ~~ pagerank(i)



Details

$$\underbrace{\sum_{i\in\mathcal{S}} 2\lambda\mathbf{u}(i)^2}_{\text{P1: \textit{node importance}}} - \underbrace{\sum_{i,j\in\mathcal{S}} \mathbf{A}(i,j)\mathbf{u}(i)\mathbf{u}(j)}_{\text{P2: \textit{set diversity}}}$$



Original Graph → Select by P1 → Select by P1+P2

Our Solution: Part 2: NetShield

- We prove that:

  $\widehat{SV}(S)$ is sub-modular (& monotone non-decreasing)

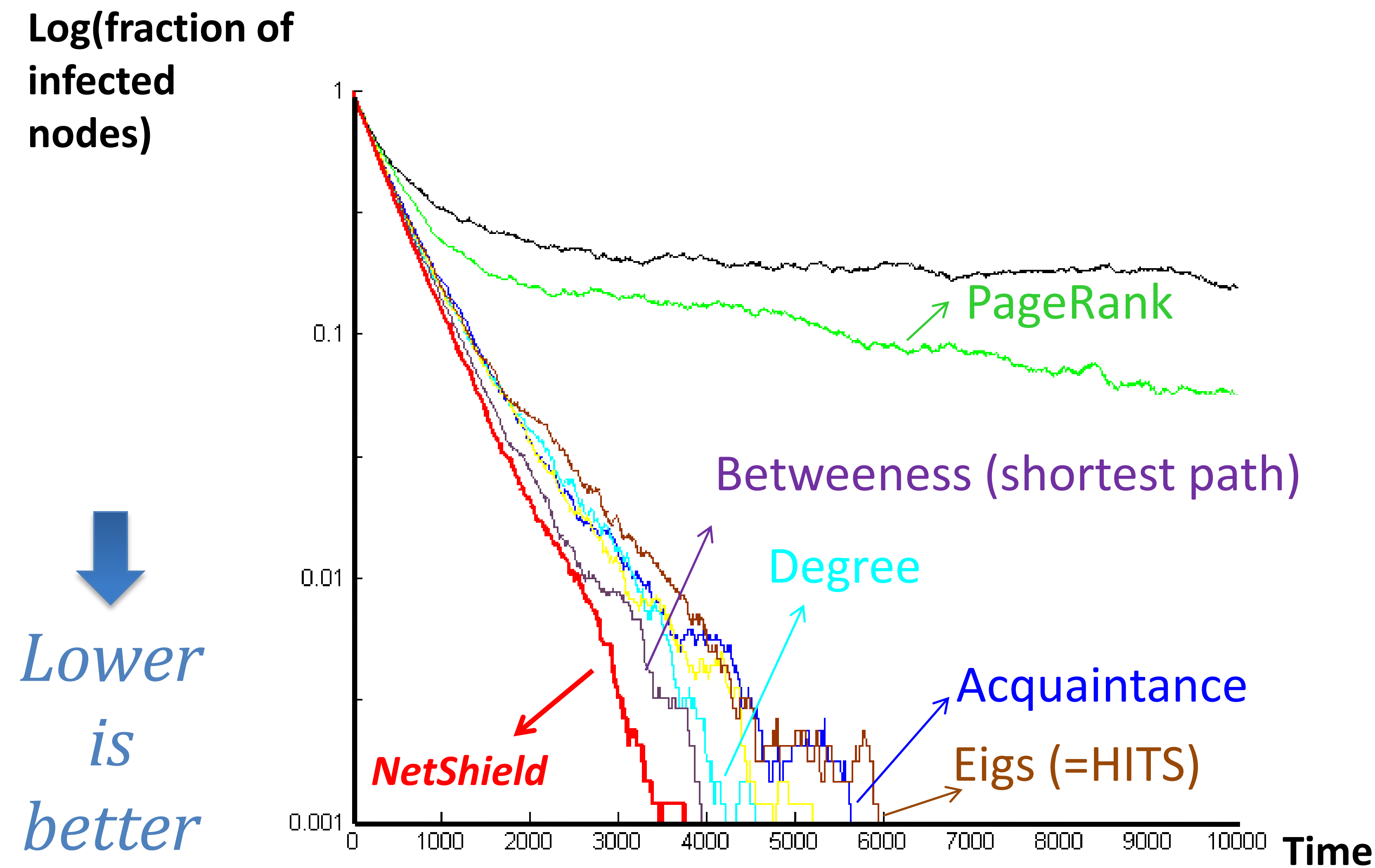
Corollary: Greedy algorithm works

1. NetShield is near-optimal (w.r.t. max $\widehat{SV}(S)$)
2. NetShield is $O(nk^2+m)$

- NetShield: Greedily add best node at each step

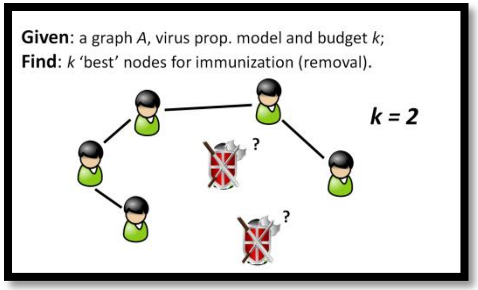
Footnote: near-optimal means $\widehat{SV}(S^{NetShield}) >= (1-1/e)\ \widehat{SV}(S^{Opt})$

Experiment: Immunization quality



Part 2: Algorithms

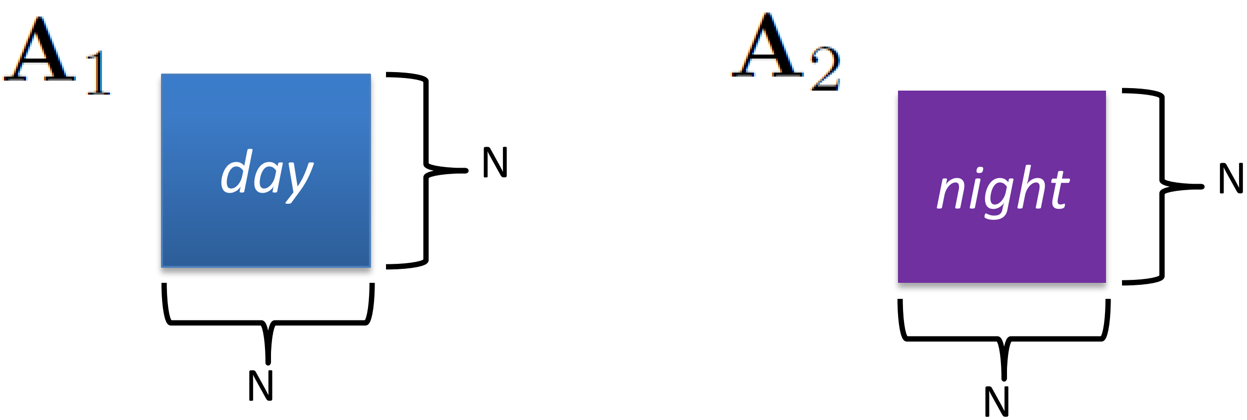
- **Q3: Whom to immunize?**
  - Full Immunization (Static Graphs)
  - **Full Immunization (Dynamic Graphs)**
  - Fractional Immunization
- Q4: How to detect outbreaks?
- Q5: Who are the culprits?



Full Dynamic Immunization

- *Given:*

  Set of *T* **arbitrary** graphs $\{\mathbf{A}_1, \mathbf{A}_2 \; \ldots, \mathbf{A}_T\}$



, weekend…..

- *Find:*

  k ‘best’ nodes to immunize (remove)

In Prakash+ ECML-PKDD 2010

Full Dynamic Immunization

$\mathbf{A}_1$ day

$\mathbf{A}_2$ night

- Our solution
  - Recall theorem
  - Simple: reduce $\lambda_{\prod_i \mathbf{S}_i}$ (=λ)

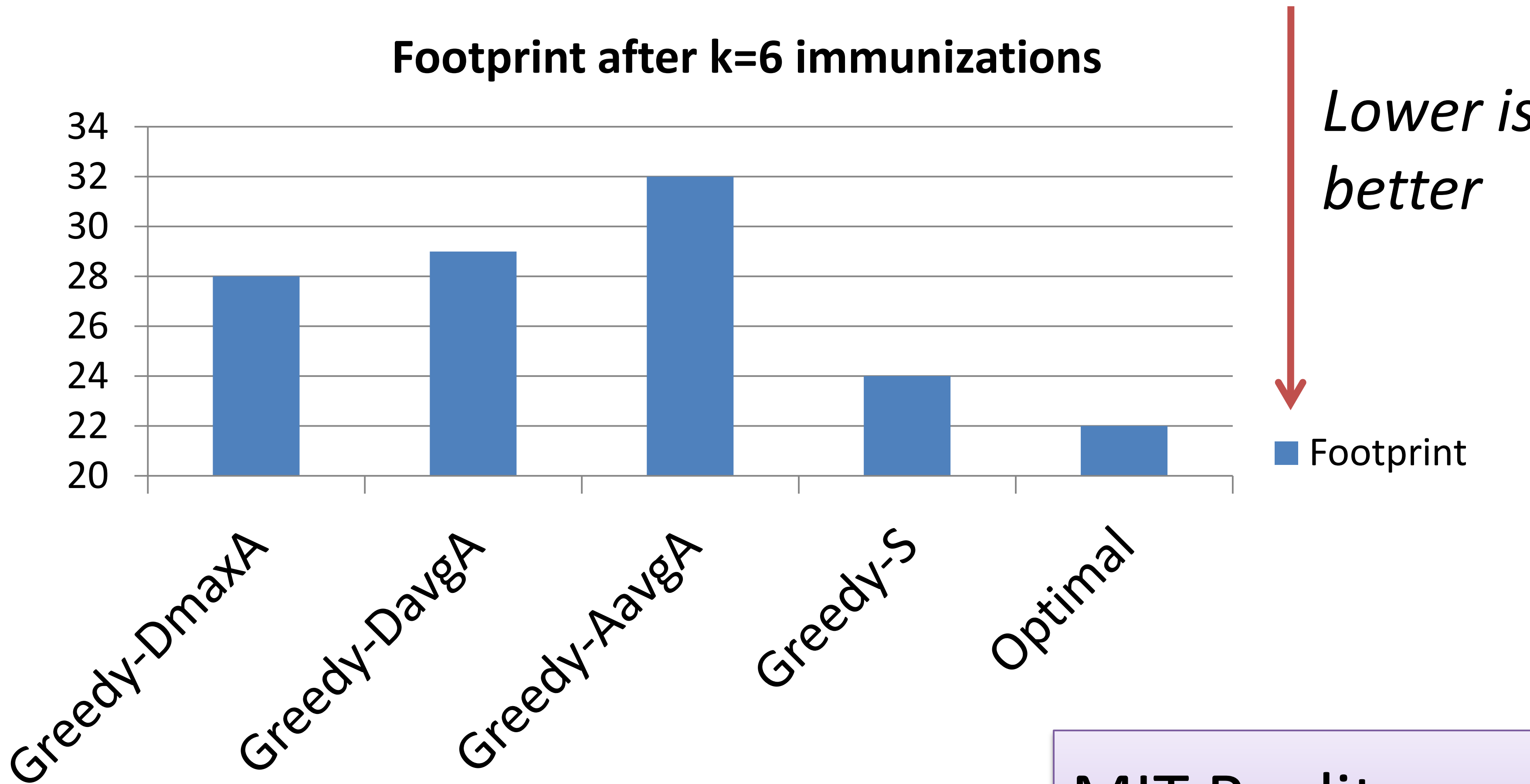
Matrix Product

- Goal: max eigendrop Δ λ

$$\Delta\lambda = \lambda_{\text{before}} - \lambda_{\text{after}}$$

- No competing policy for comparison
- We propose and evaluate many policies

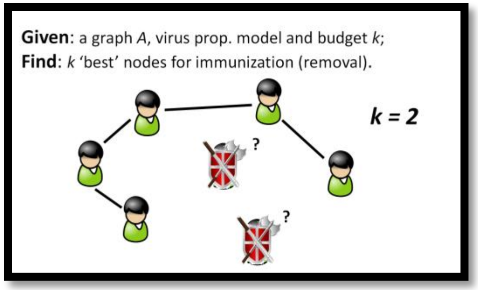
Performance of Policies



MIT Reality Mining

Part 2: Algorithms

- **Q3: Whom to immunize?**
  - Full Immunization (Static Graphs)
  - Full Immunization (Dynamic Graphs)
  - **Fractional Immunization**
- Q4: How to detect outbreaks?
- Q5: Who are the culprits?



Fractional Immunization of Networks

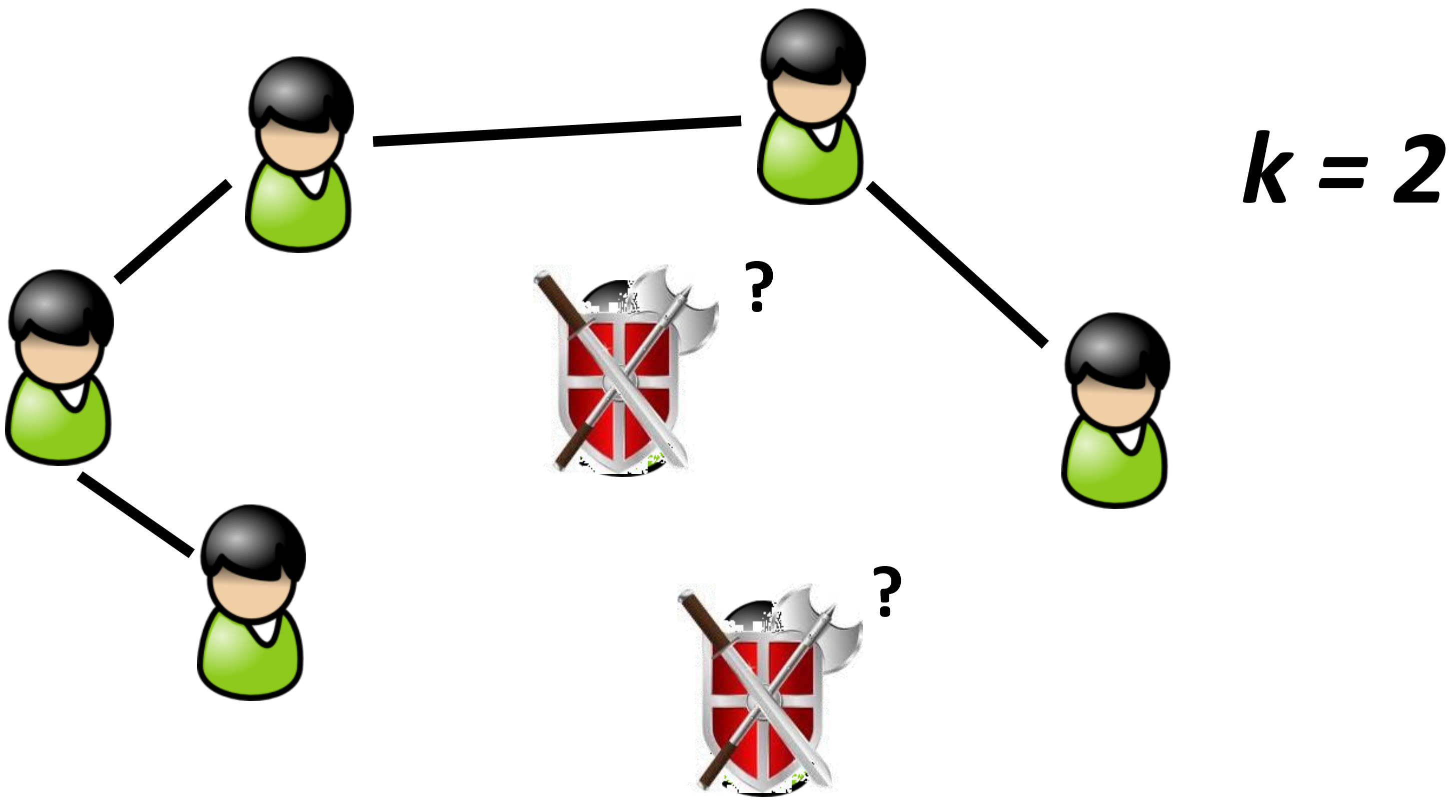
B. Aditya Prakash, Lada Adamic, Theodore Iwashyna (M.D.), Hanghang Tong, Christos Faloutsos

Under Submission

Previously: Full Static Immunization

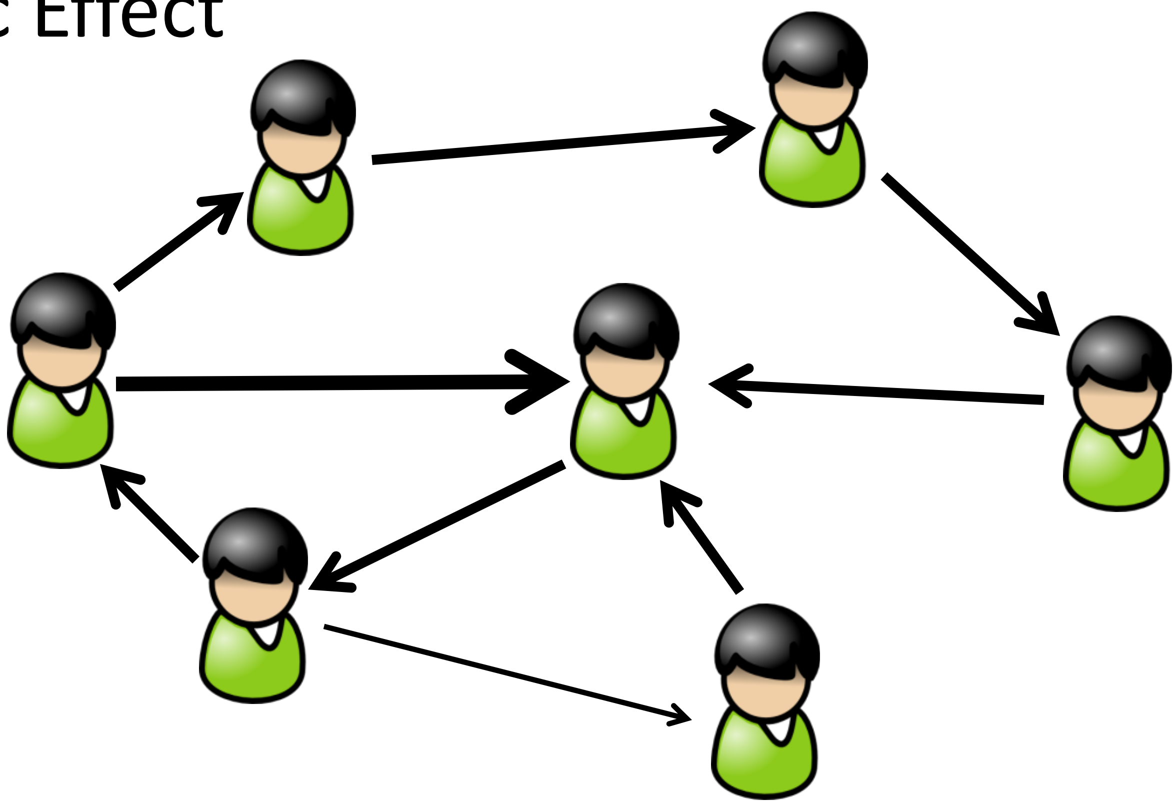
Given: a graph *A*, virus prop. model and budget *k*;

Find: *k* ‘best’ nodes for immunization (removal).



Fractional Asymmetric Immunization

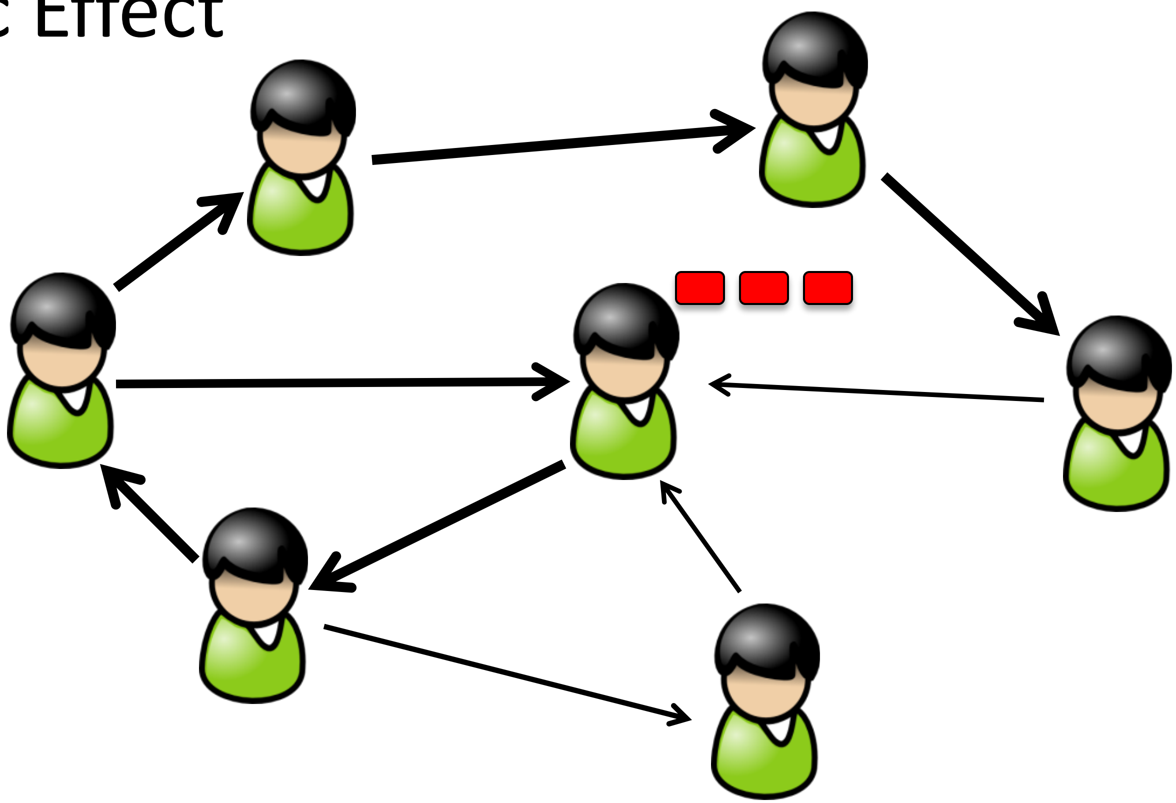
- Fractional Effect [$f(x) = 0.5^x$]
- Asymmetric Effect



 # antidotes = 3

Now: Fractional Asymmetric Immunization

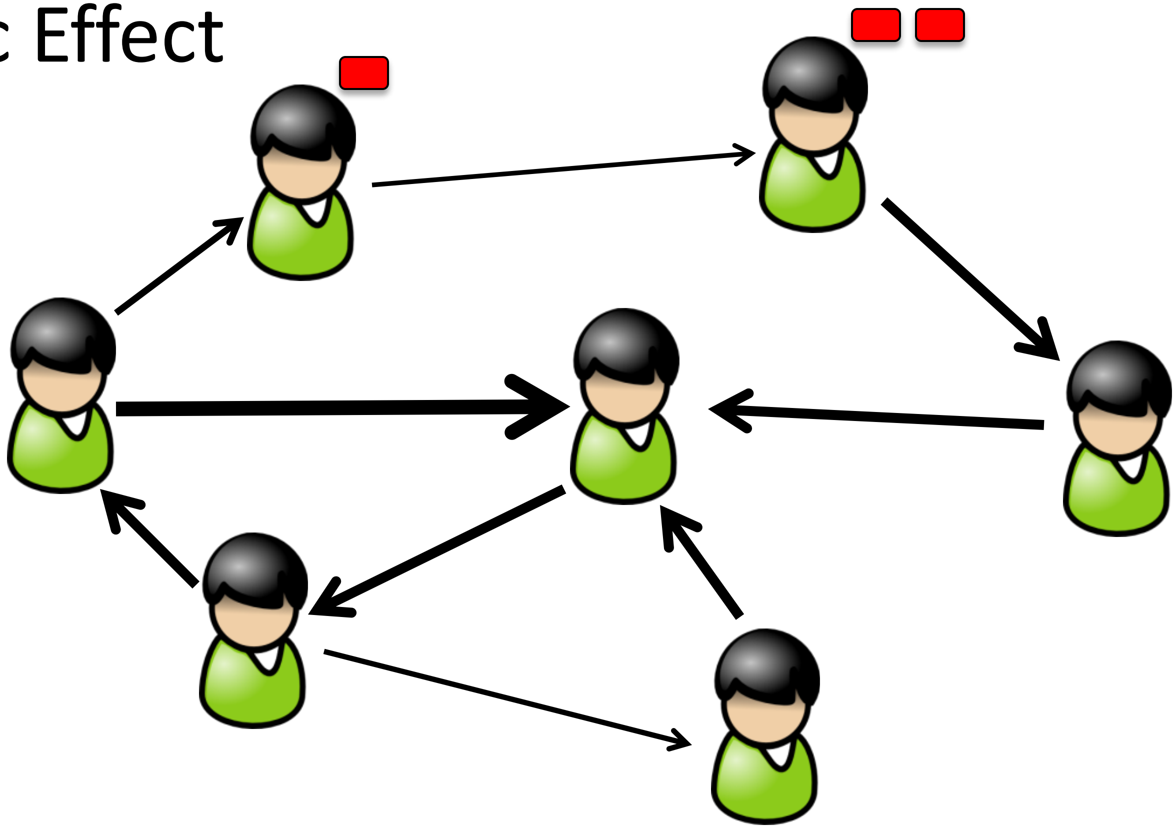
- Fractional Effect [$f(x) = 0.5^x$]
- Asymmetric Effect



 # antidotes = 3

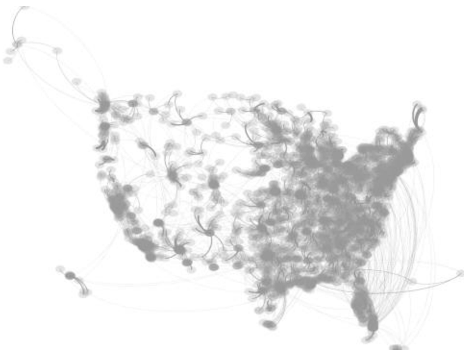
Fractional Asymmetric Immunization

- Fractional Effect [$f(x) = 0.5^x$]
- Asymmetric Effect

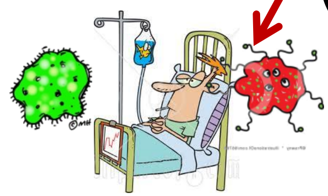


 # antidotes = 3

Fractional Asymmetric Immunization



Drug-resistant Bacteria (like XDR-TB)

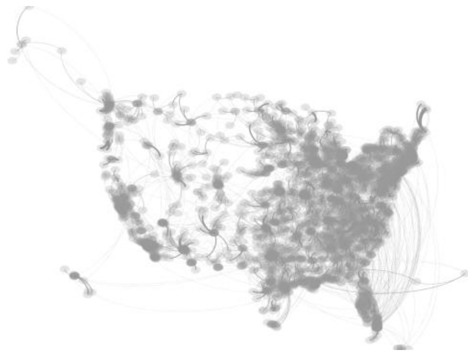


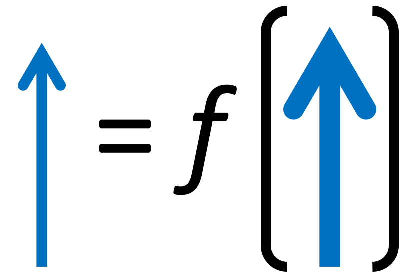
Hospital

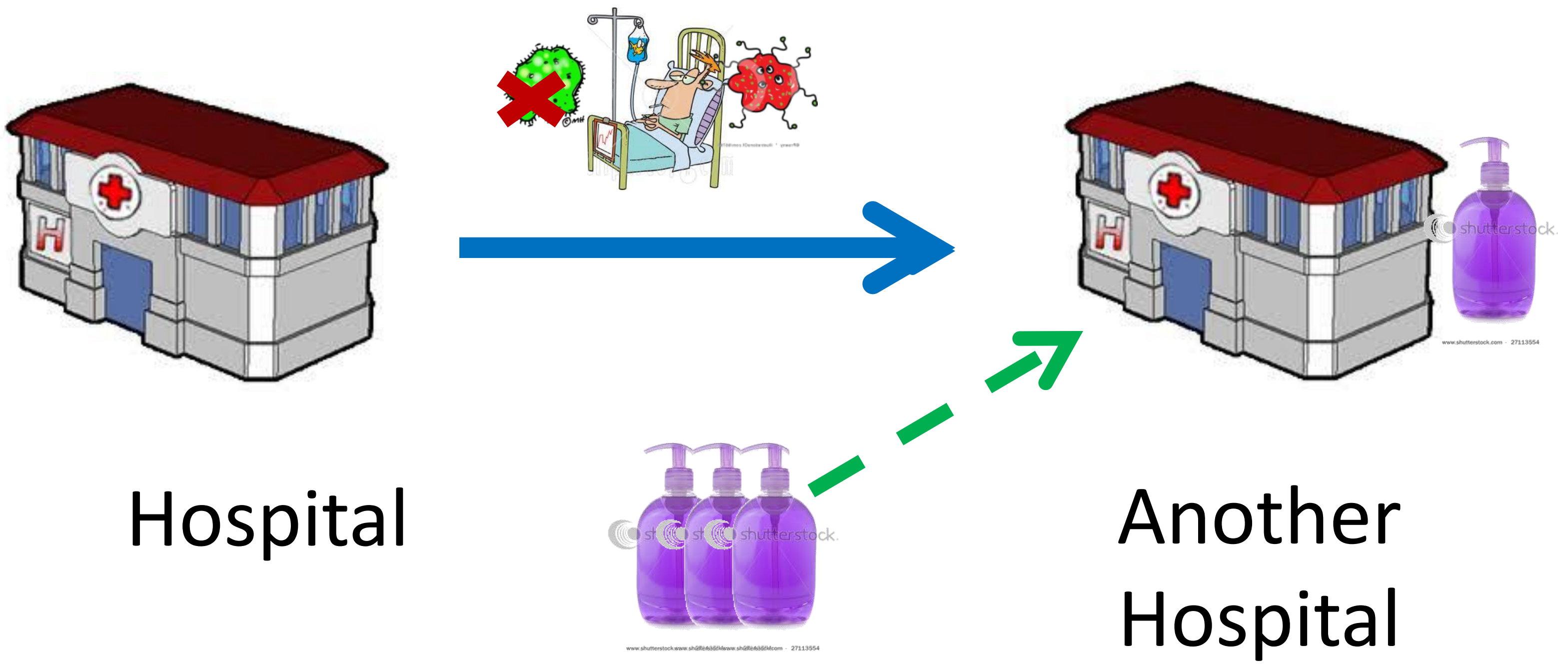
Another Hospital

Prakash and Faloutsos 2012

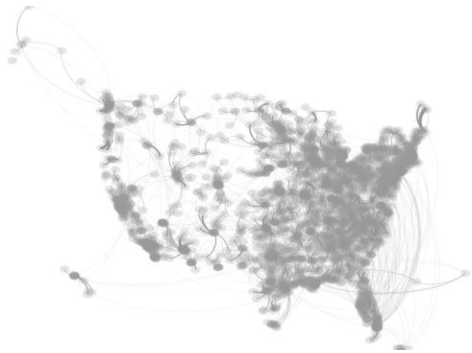
Fractional Asymmetric Immunization



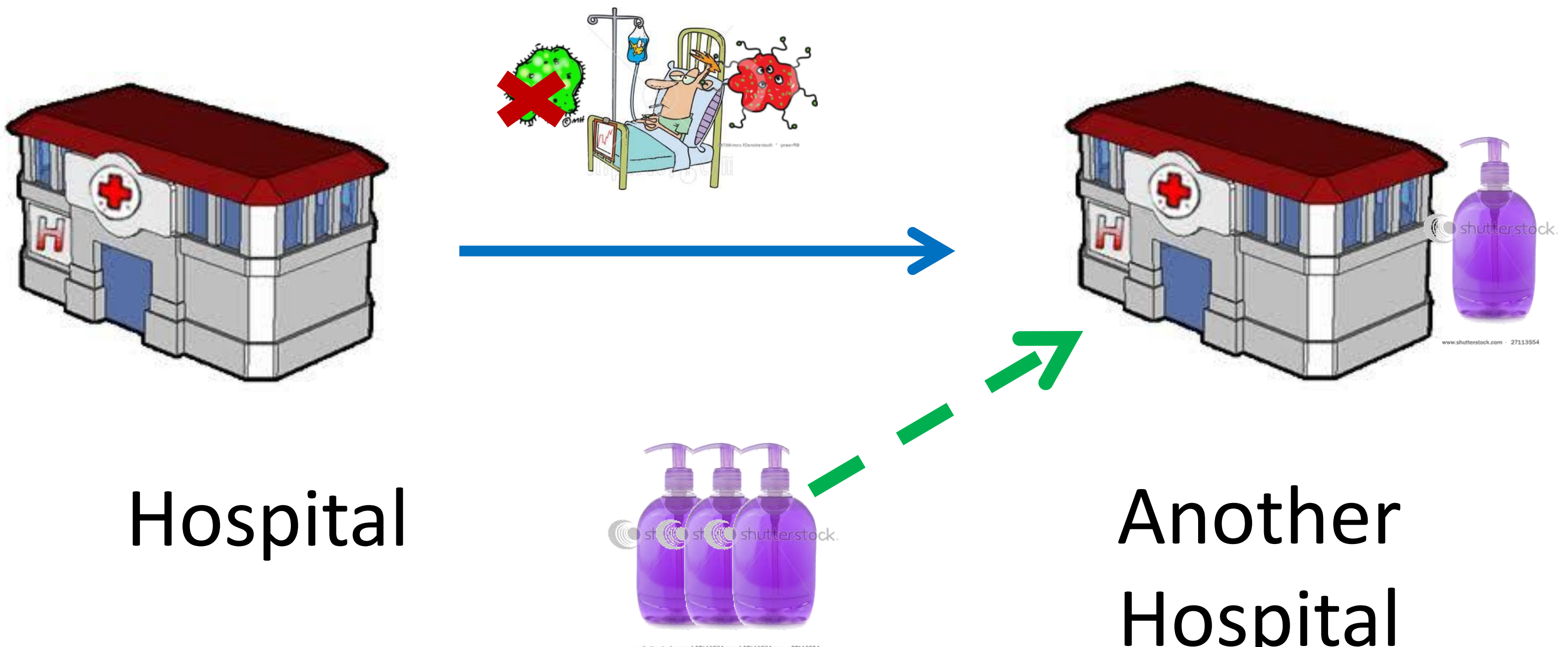




Fractional Asymmetric Immunization



***Problem**: Given k units of disinfectant, how to distribute them to maximize hospitals saved?*

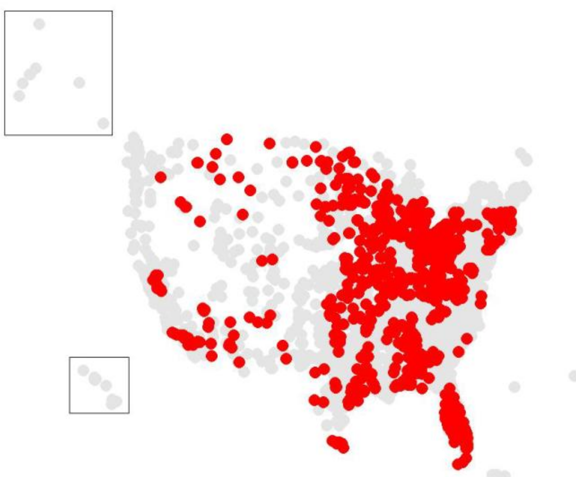


Prakash and Faloutsos 2012

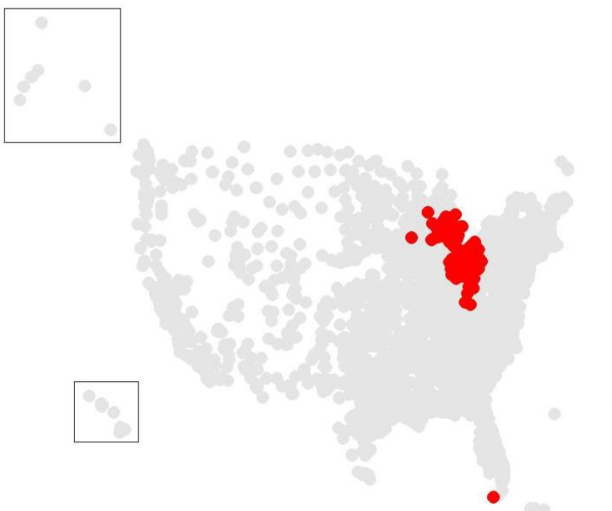
Our Algorithm "SMART-ALLOC"

[US-MEDICARE NETWORK 2005]

- Each circle is a hospital, ~3000 hospitals
- More than 30,000 patients transferred

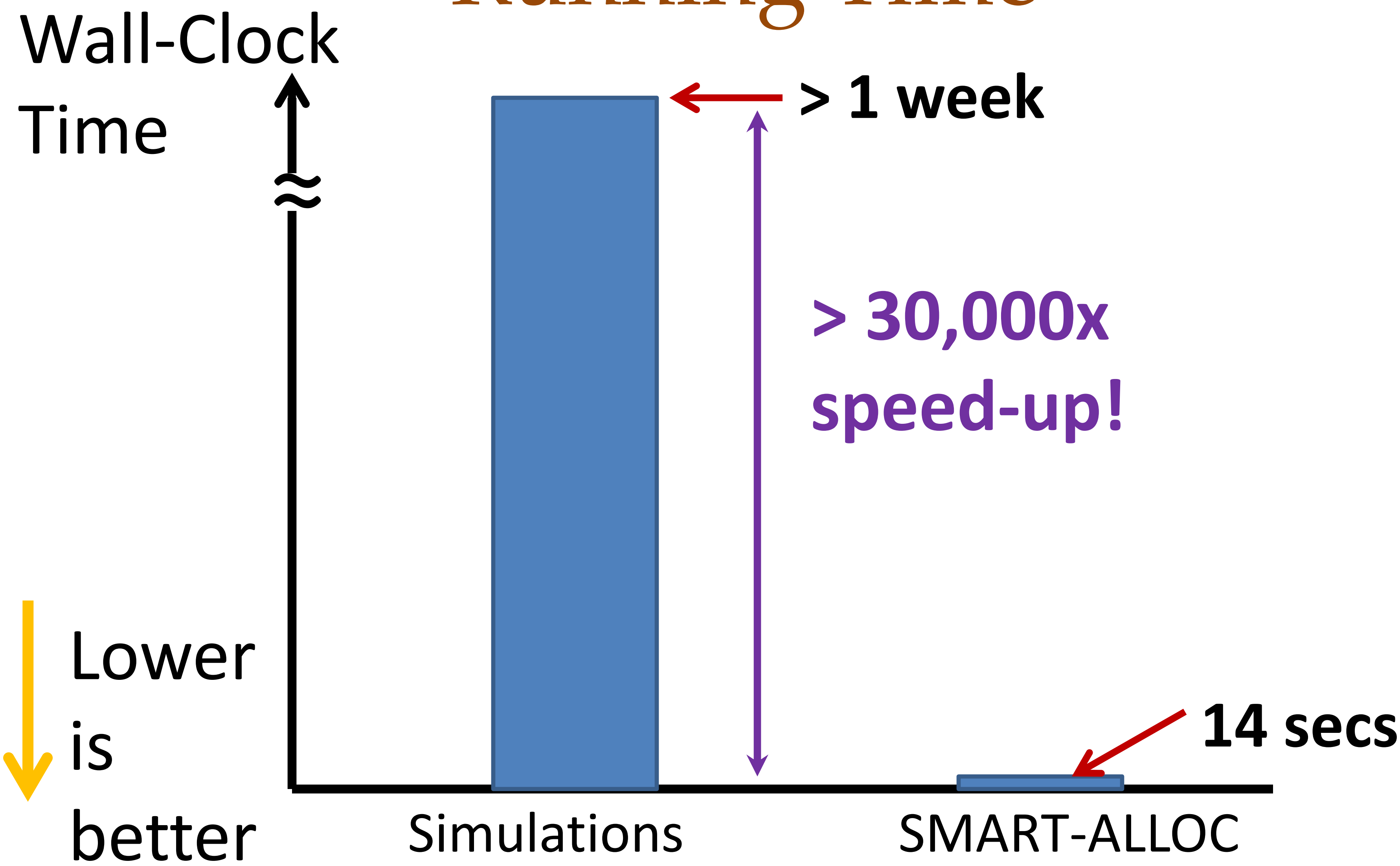


CURRENT PRACTICE



SMART-ALLOC

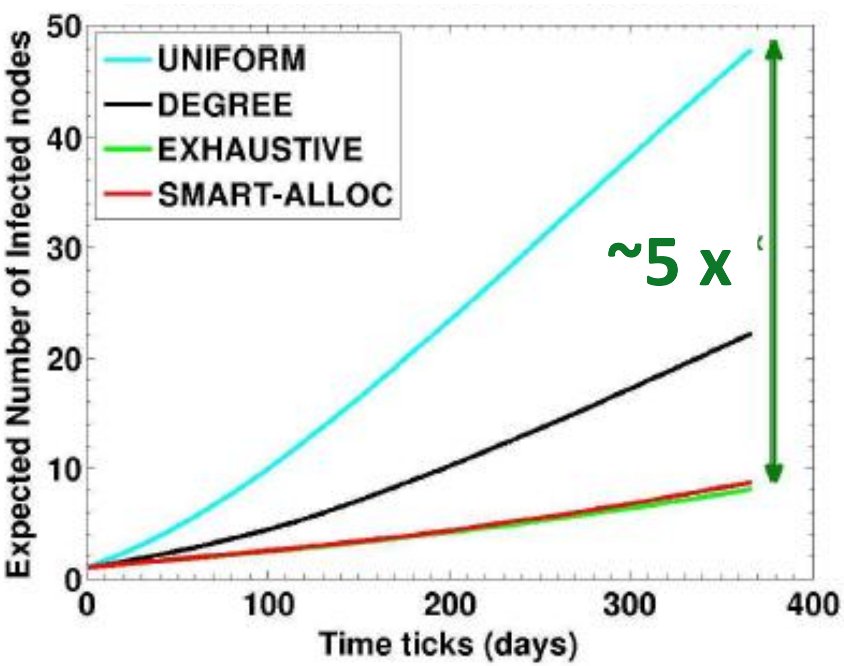
Running Time



Lower is better

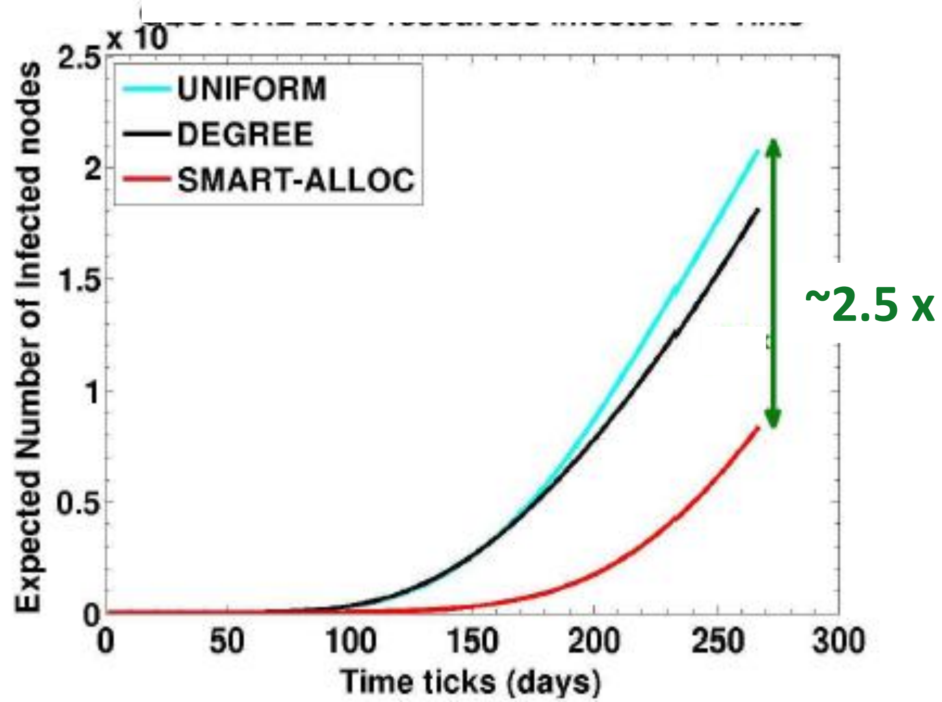
Experiments

PENN-NETWORK



K = 200

SECOND-LIFE



K = 2000

Part 2: Algorithms

- Q3: Whom to immunize?
- **Q4: How to detect outbreaks?**
- Q5: Who are the culprits?

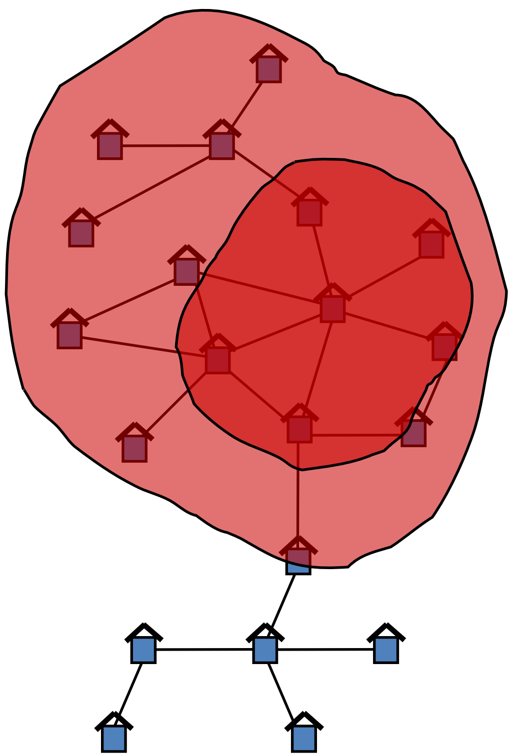
Break!

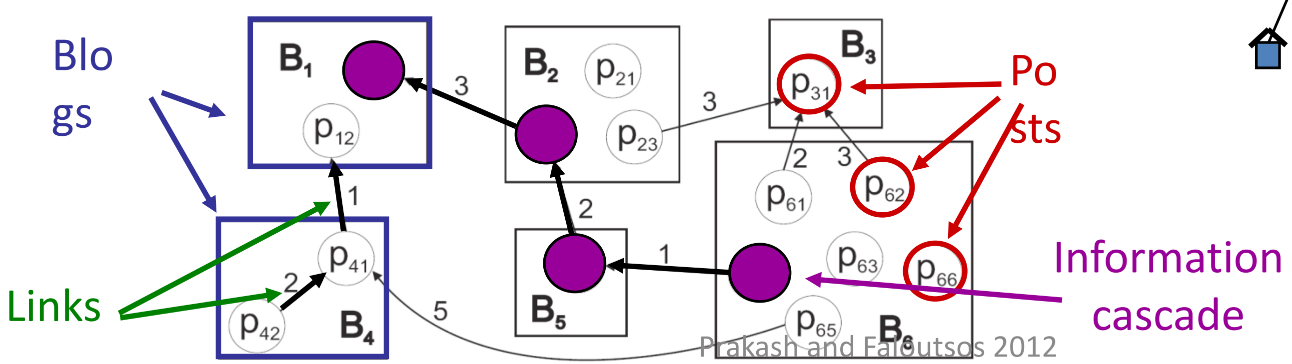
Part 2: Algorithms

- Q3: Whom to immunize?
- **Q4: How to detect outbreaks?**
- Q5: Who are the culprits?

Outbreak detection

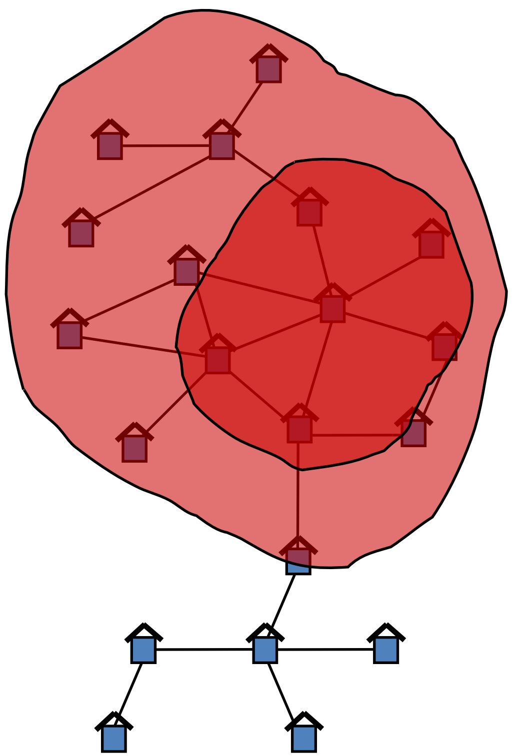
- Spot contamination points
  - Minimize time to detection, population affected
  - Maximize probability of detection.
  - Minimize sensor placement cost.

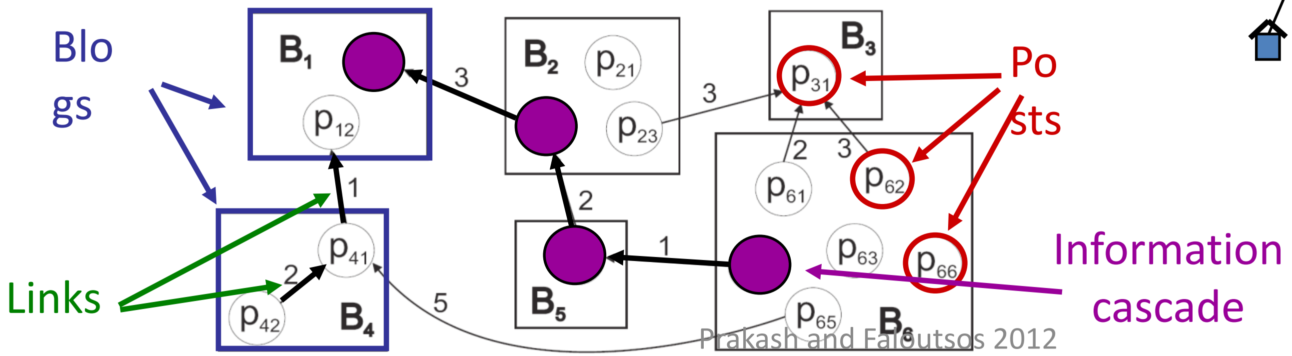




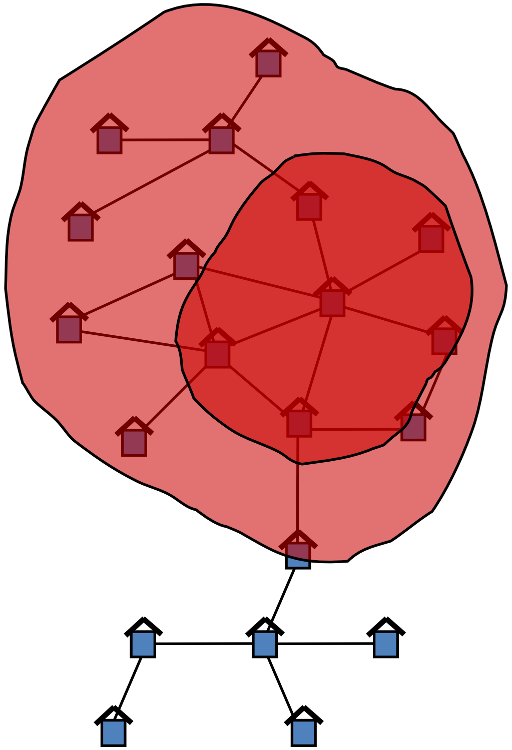
Outbreak detection

- Spot `hot blogs'
  - Minimize time to detection, population affected
  - Maximize probability of detection.
  - Minimize sensor placement cost.





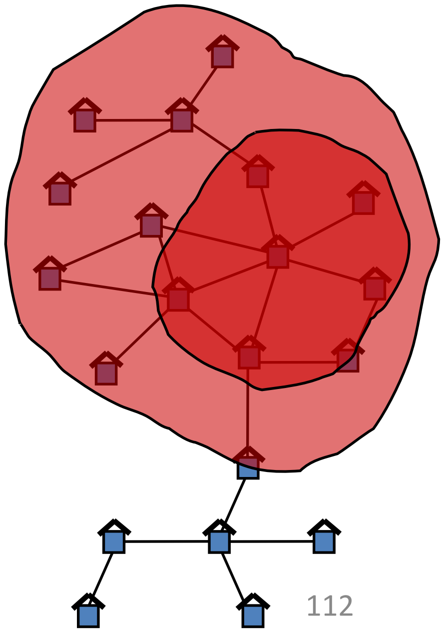
- J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, N. Glance. "Cost-effective Outbreak Detection in Networks" KDD 2007



CELF: Main idea

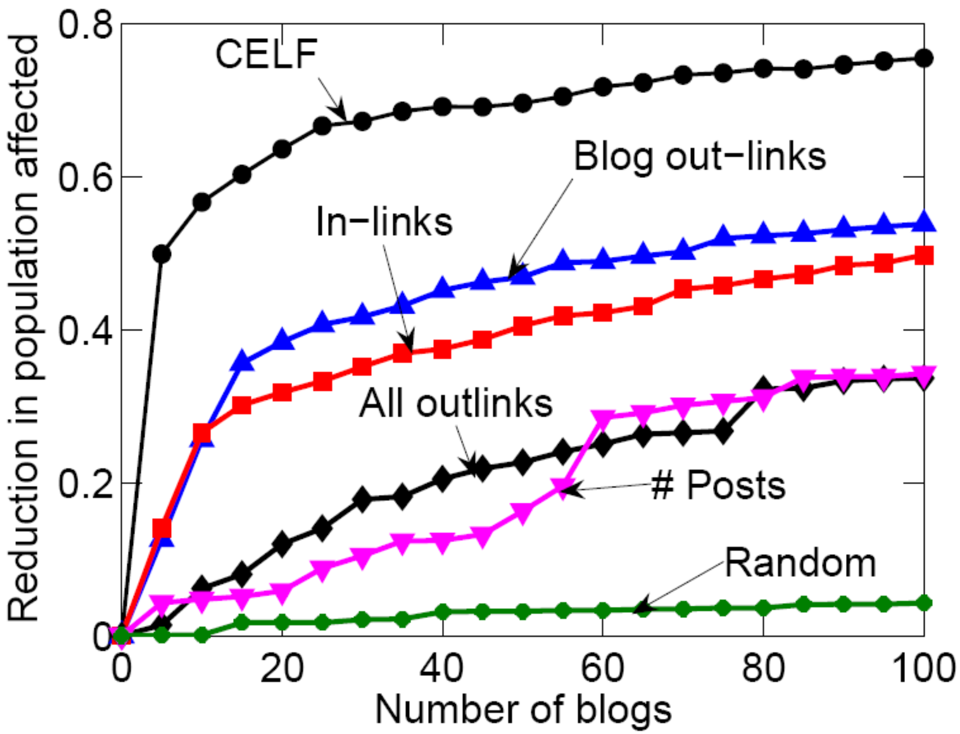
- Given: a graph $G(V,E)$
  - a budget of B sensors
  - data on how contaminations spread over the network:
- Place the sensors
- To minimize time to detect outbreak

CELF algorithm uses submodularity and lazy evaluation



Blogs: Comparison to heuristics

Benefit
(higher=
better)



"Best 10 blogs to read"

NP - number of posts, IL- in-links, OLO- blog out links, OLA- all out links

k	PA score	Blog	NP	IL	OLO	OLA
1	0.1283	http://instapundit.com	4593	4636	1890	5255
2	0.1822	http://donsurber.blogspot.com	1534	1206	679	3495
3	0.2224	http://sciencepolitics.blogspot.com	924	576	888	2701
4	0.2592	http://www.watcherofweasels.com	261	941	1733	3630
5	0.2923	http://michellemalkin.com	1839	12642	1179	6323
6	0.3152	http://blogometer.nationaljournal.com	189	2313	3669	9272
7	0.3353	http://themodulator.org	475	717	1844	4944
8	0.3508	http://www.bloggersblog.com	895	247	1244	10201
9	0.3654	http://www.boingboing.net	5776	6337	1024	6183
10	0.3778	http://atrios.blogspot.com	4682	3205	795	3102

Part 2: Algorithms

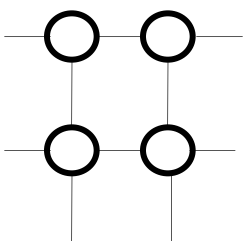
- Q3: Whom to immunize?
- Q4: How to detect outbreaks?
- **Q5: Who are the culprits?**

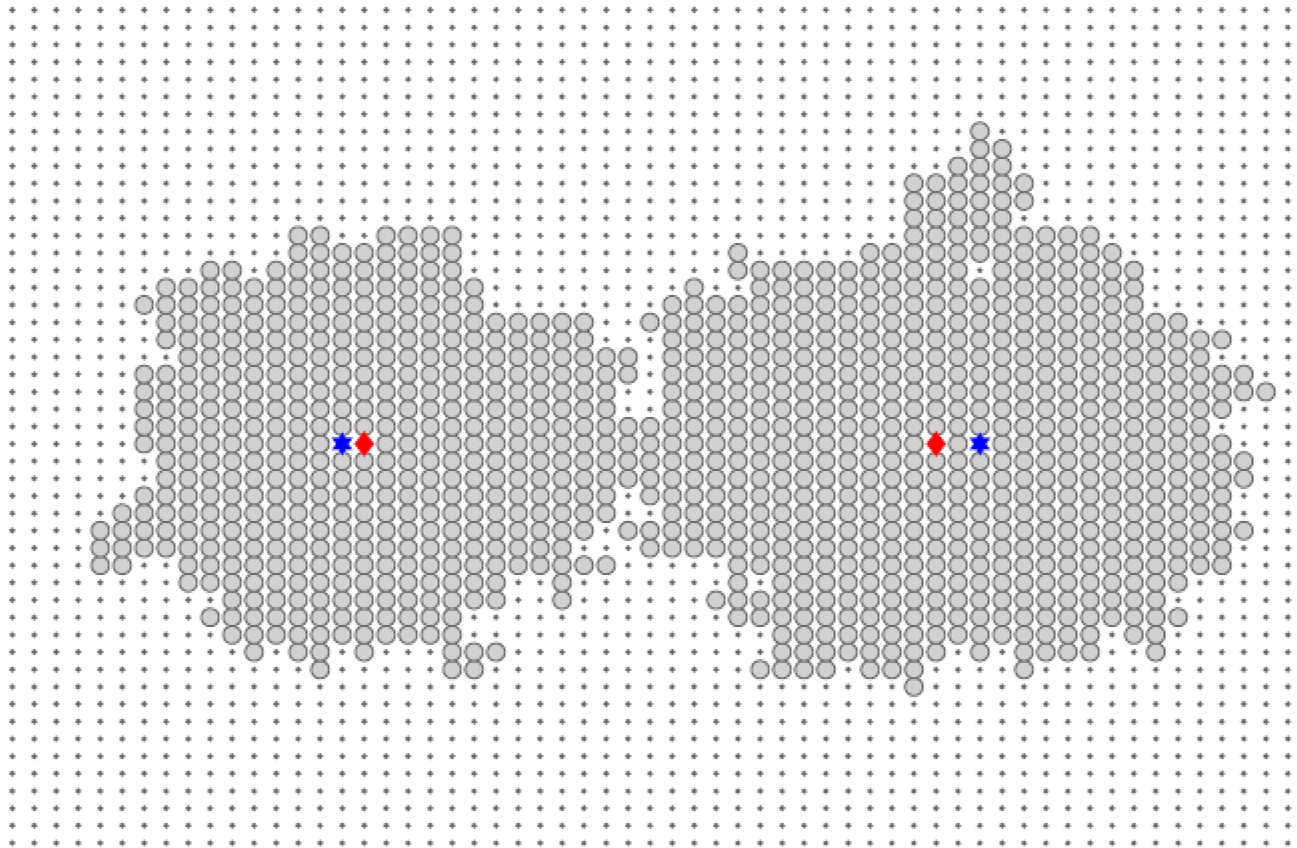
- B. Aditya Prakash, Jilles Vreeken, Christos Faloutsos ‘Detecting Culprits in Epidemics: Who and How many?’

  ICDM 2012, Brussels

Problem definition

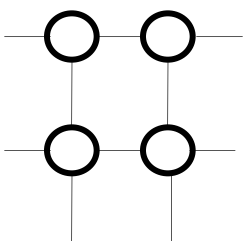
2-d grid
'+' -> infected
Who started it?

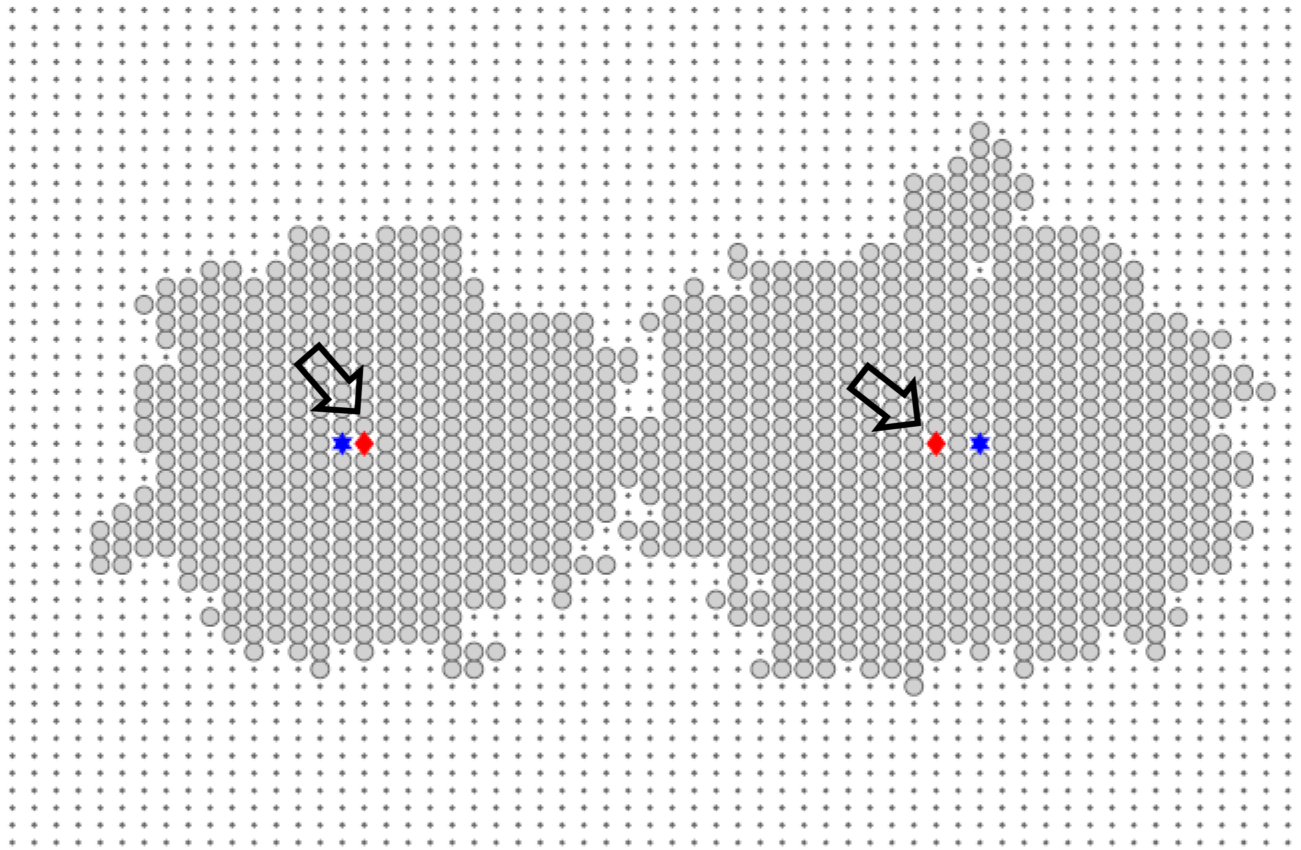




Problem definition

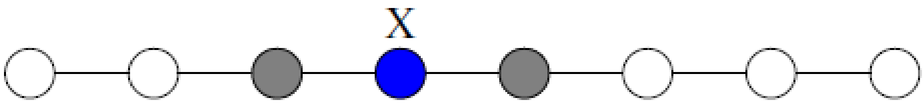
2-d grid
'+' -> infected
Who started it?





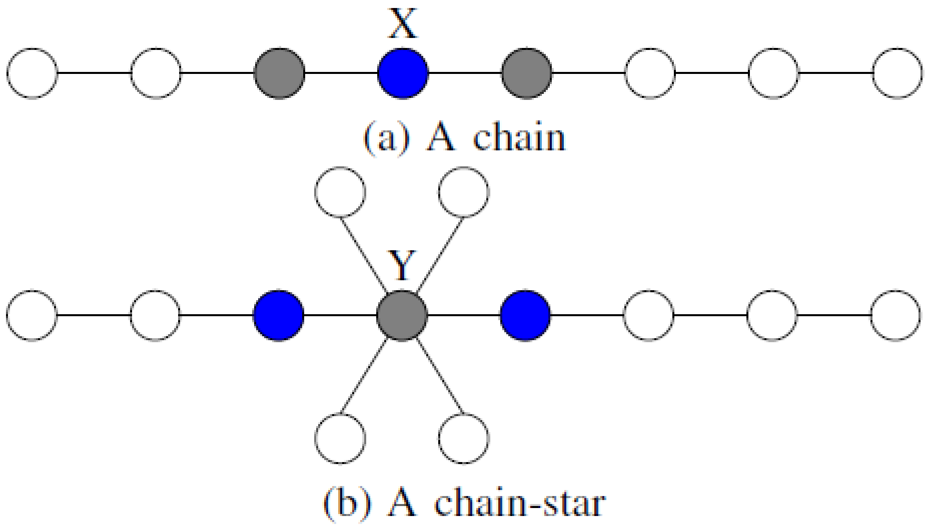
Prior work:
[Lappas et al. 2010, Shah et al. 2011]

Culprits: Exoneration

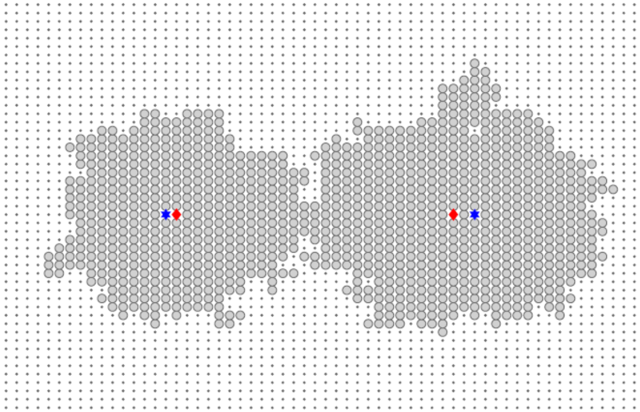


(a) A chain

Culprits: Exoneration



Who are the culprits



- Two-part solution
  - use MDL for *number* of seeds
  - for a given number:
    - exoneration = centrality + penalty
    - our method uses *smallest* eigenvector of Laplacian submatrix

- Running time = $O(k^*(\mathcal{E}_I + \mathcal{E}_F + \mathcal{V}_I))$
  - linear! (in edges and nodes)

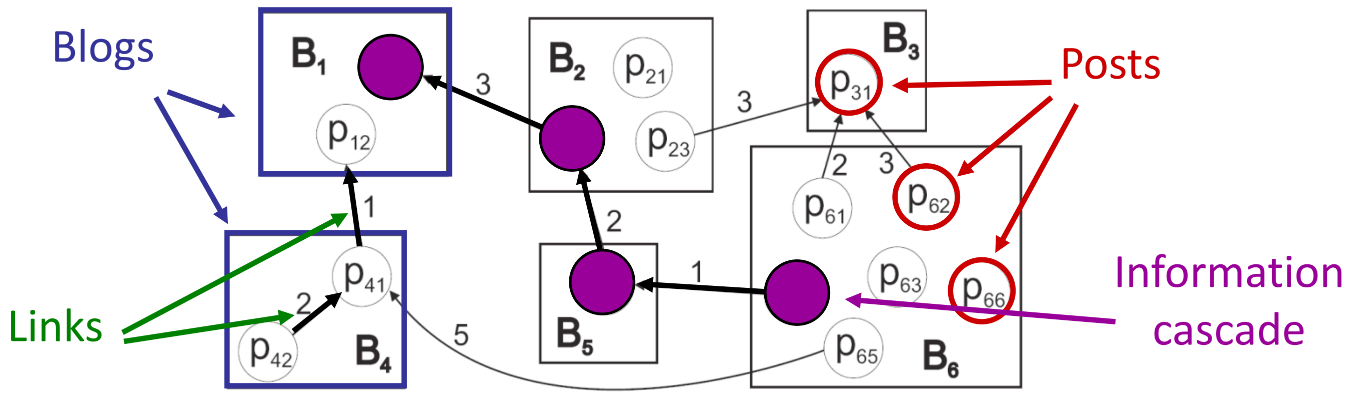
Outline

- Motivation
- Part 1: Understanding Epidemics (Theory)
- Part 2: Policy and Action (Algorithms)
- **Part 3: Learning Models (Empirical Studies)**
- Conclusion

Part 3: Empirical Studies

- **Q6: How do cascades look like?**
- Q7: How does activity evolve over time?
- Q8: How does external influence act?

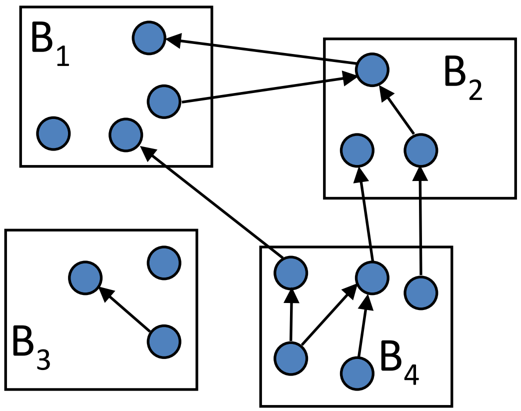
Cascading Behavior in Large Blog Graphs



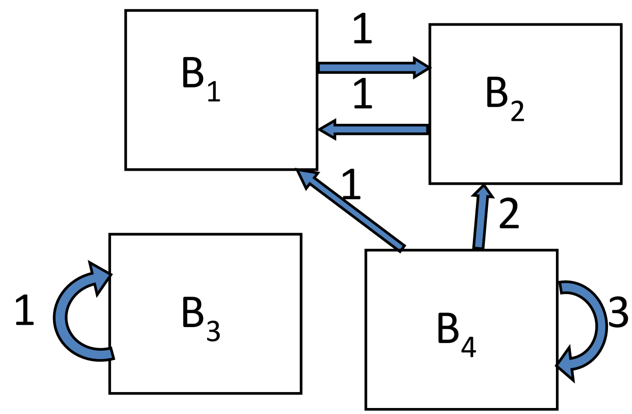
How does information propagate over the blogosphere?

J. Leskovec, M.McGlohon, C. Faloutsos, N. Glance, M. Hurst. Cascading Behavior in Large Blog Graphs. SDM 2007.

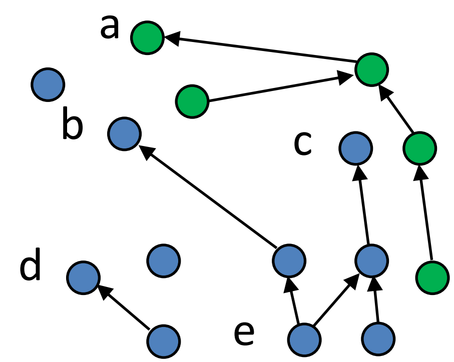
Cascades on the Blogosphere



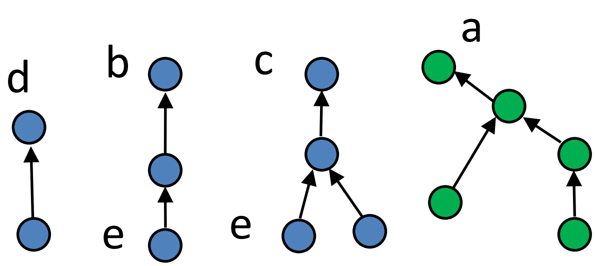
Blogosphere
blogs + posts



Blog network
links among blogs



Post network
links among posts

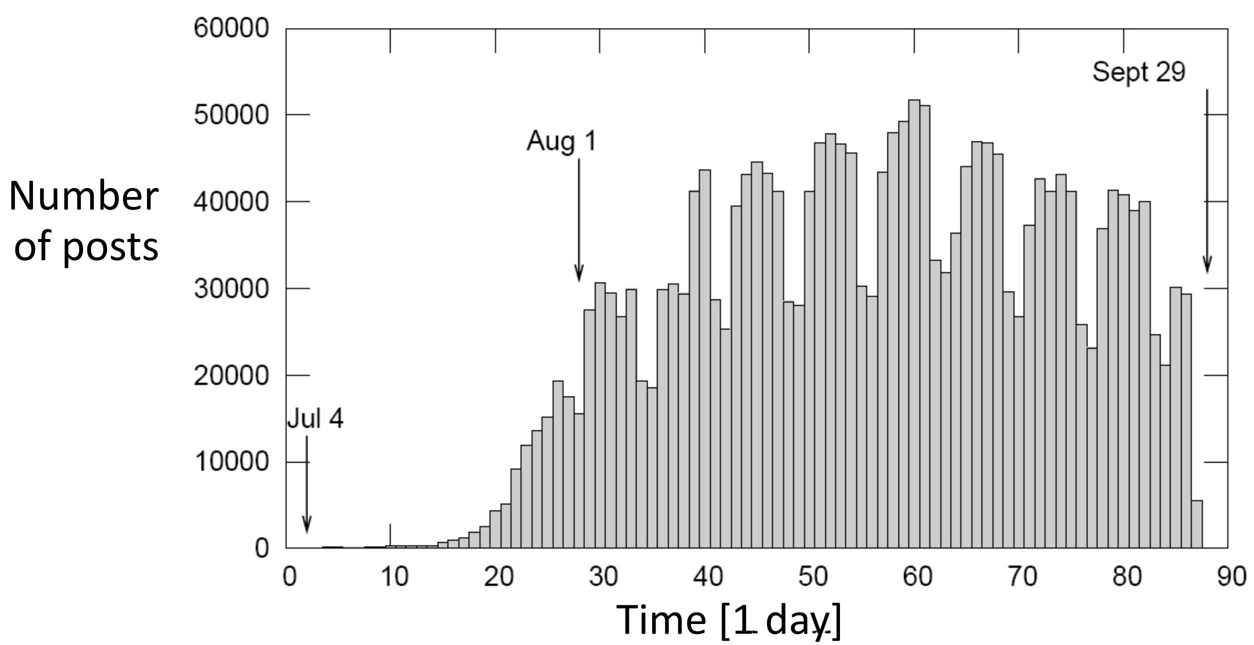


Cascades

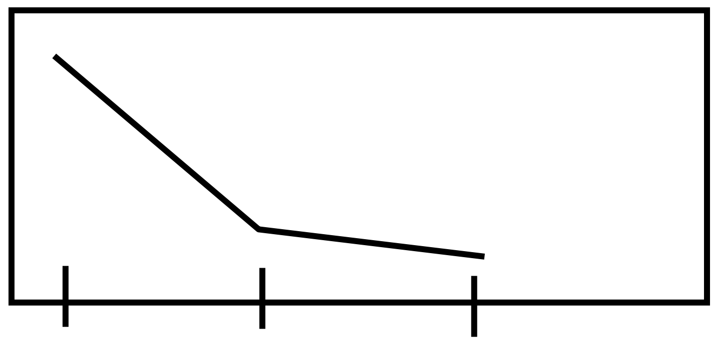
Cascade is graph induced by a time ordered propagation of information (edges)

Blog data

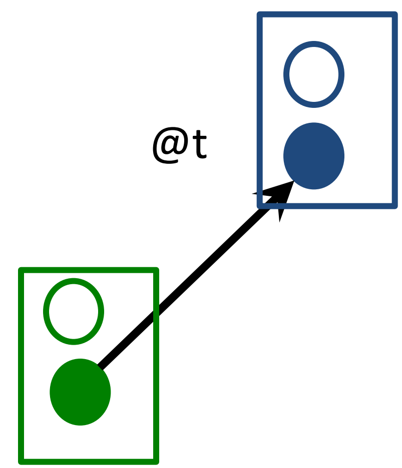
- 45,000 blogs participating in cascades
- All their posts for 3 months (Aug-Sept '05)
- 2.4 million posts
- ~5 million links (245,404 inside the dataset)



Popularity over time

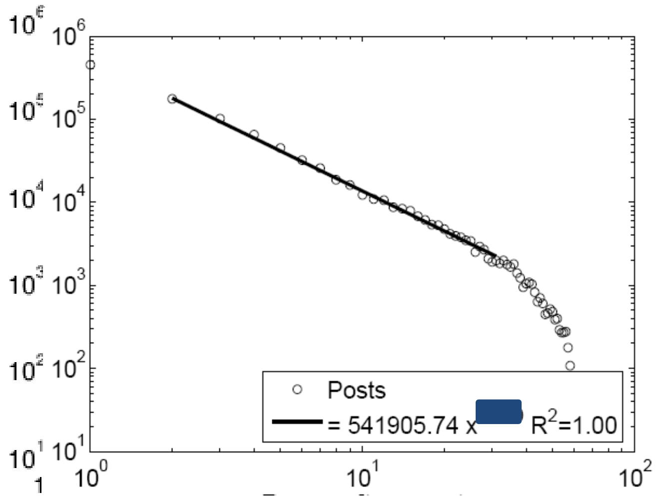


Post popularity drops-off – exponentially?



Popularity over time

in links
(**log**)

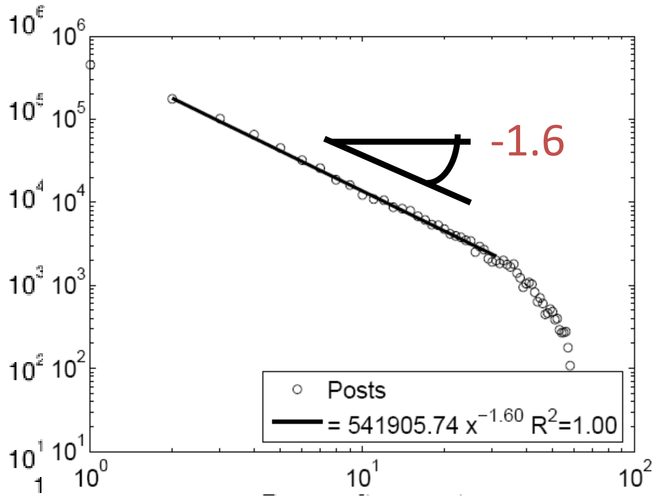


days after post
(**log**)

Post popularity drops-off – exponentially?
POWER LAW!
Exponent?

Popularity over time

in links
(**log**)



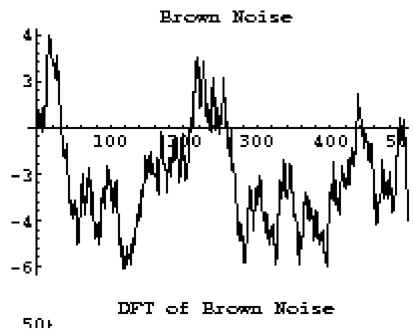
days after post
(**log**)

Post popularity drops-off – exponentially?

POWER LAW!

Exponent? -1.6

- close to -1.5: Barabasi's stack model
- and like the zero-crossings of a random walk



-1.5 slope

J. G. Oliveira & A.-L. Barabási Human Dynamics: The Correspondence Patterns of Darwin and Einstein. *Nature* **437,** 1251 (2005) . [PDF]

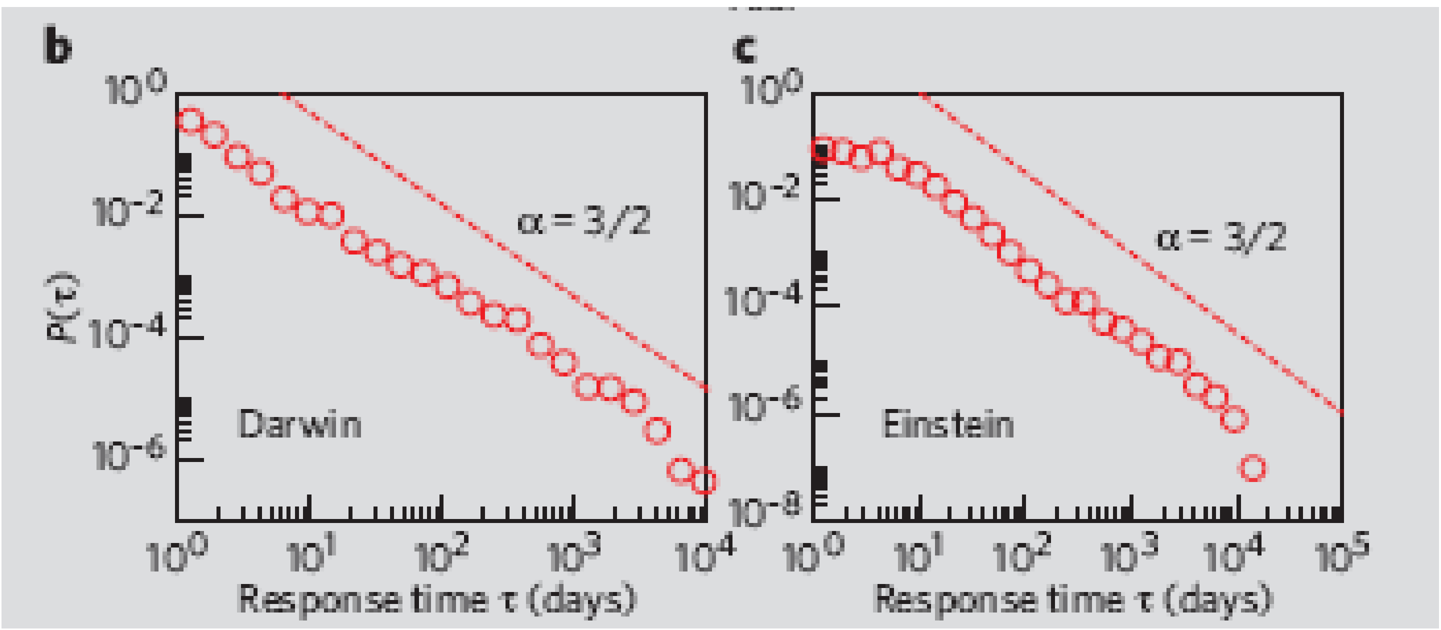
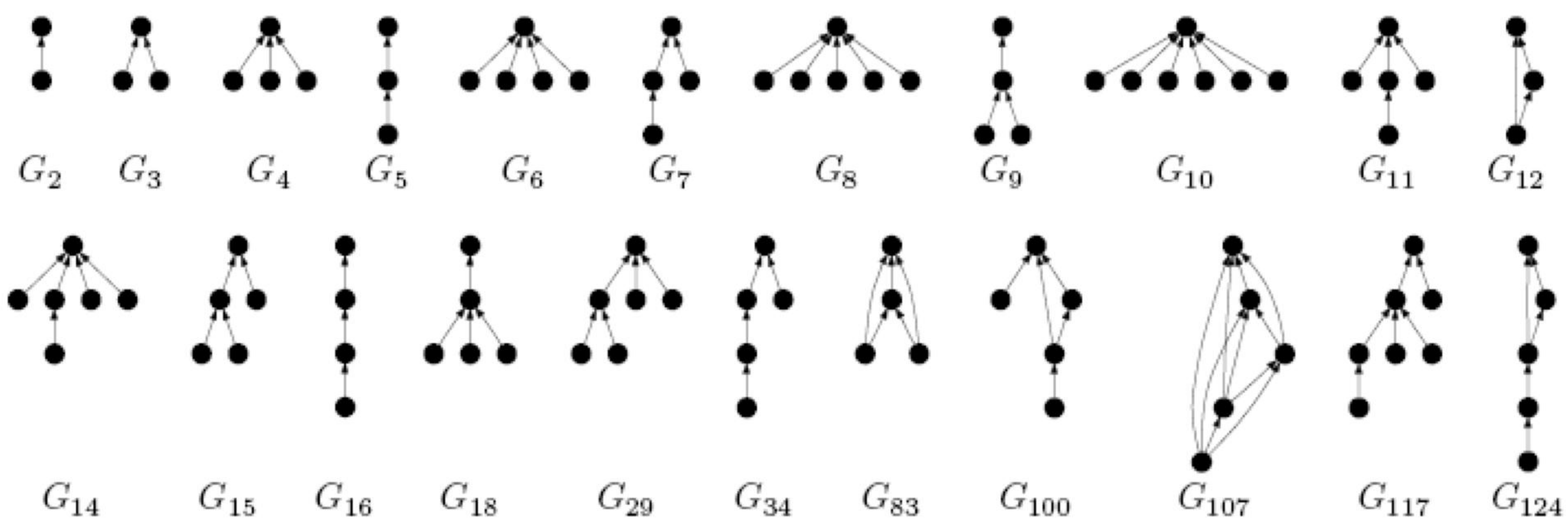


Figure 1 | The correspondence patterns of Darwin and Einstein.

Topological Observations

How do we measure how information flows through the network?

Common cascade shapes extracted using algorithms in [Leskovec, Singh, Kleinberg; PAKDD 2006].



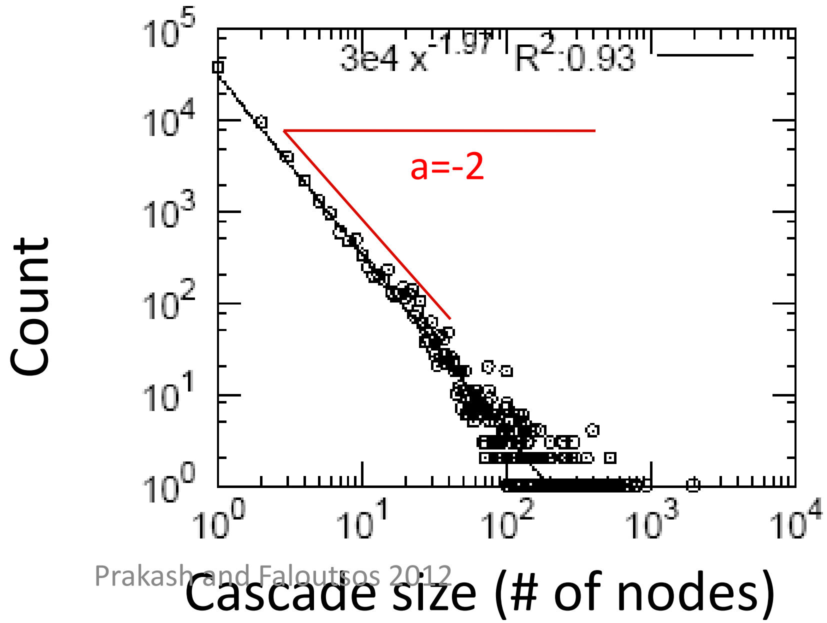
Topological Observations

What graph properties do cascades exhibit?

Cascade size distributions also follow power law.

Observation 2: *The probability of observing a cascade on n nodes follows a Zipf distribution:*

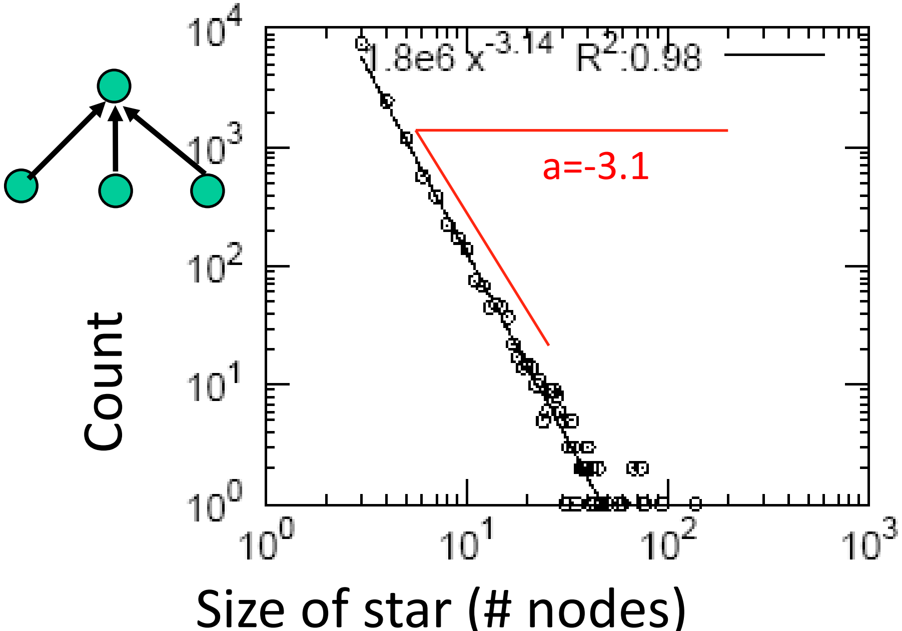
$p(n) \propto n^{-2}$

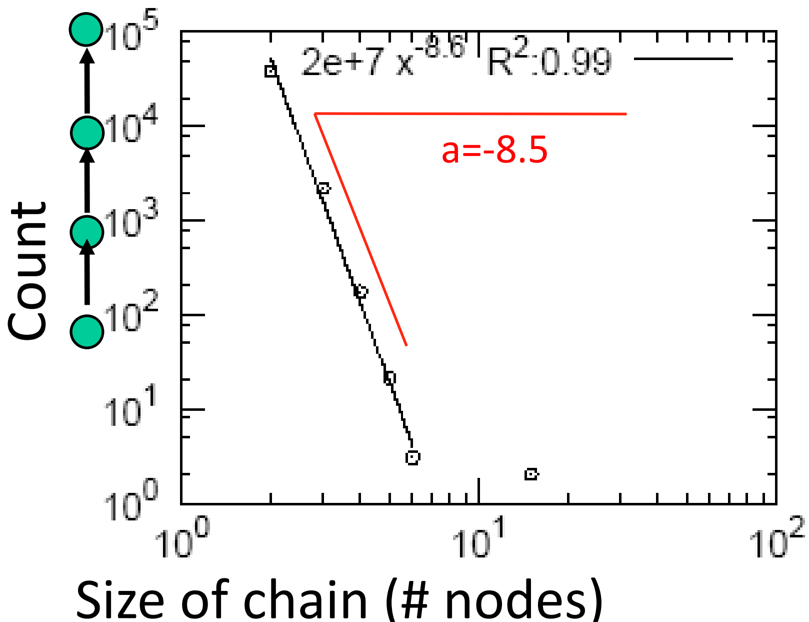


Topological Observations

What graph properties do cascades exhibit?

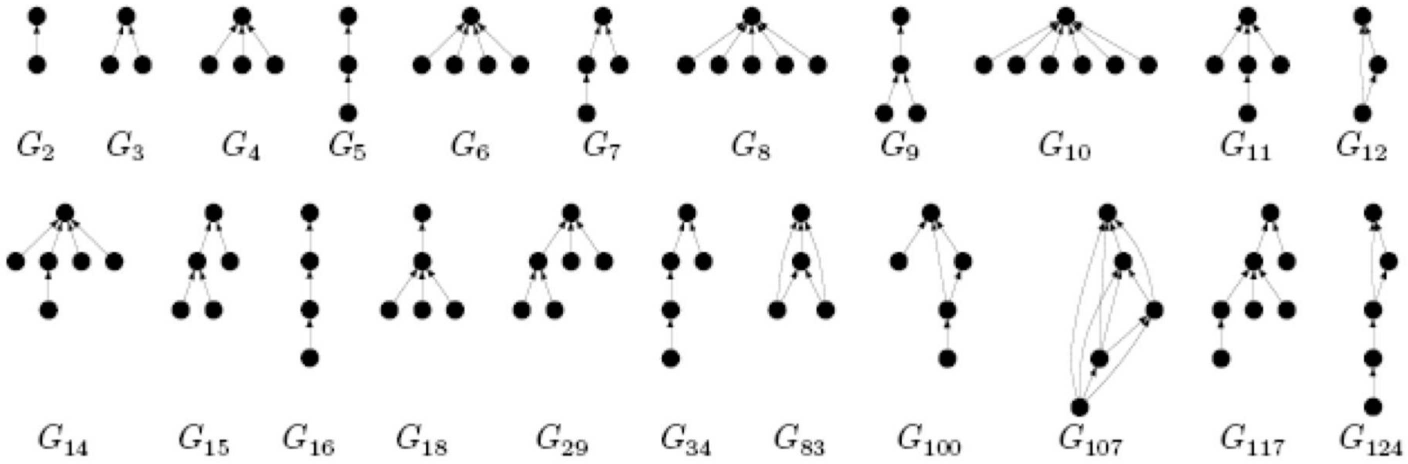
Stars and chains also follow a power law, with different exponents (star -3.1, chain -8.5).





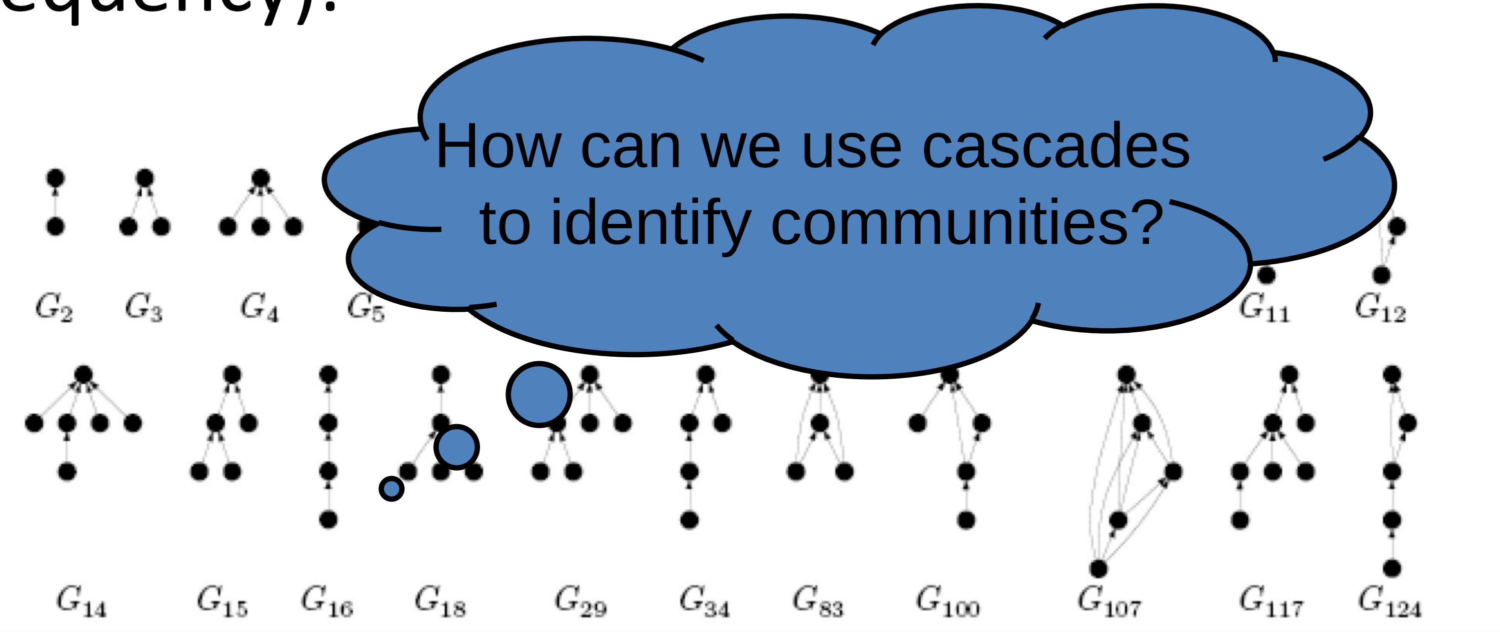
Blogs and structure

- Cascades take on different shapes (sorted by frequency):



Blogs and structure

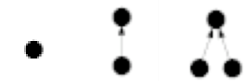
- Cascades take on different shapes (sorted by frequency):



PCA on cascade types

- Perform PCA on sparse matrix.
- Use log(count+1)
- Project onto 2 PC…

~9,000 cascade types

% ->  …………

~44,000 blogs

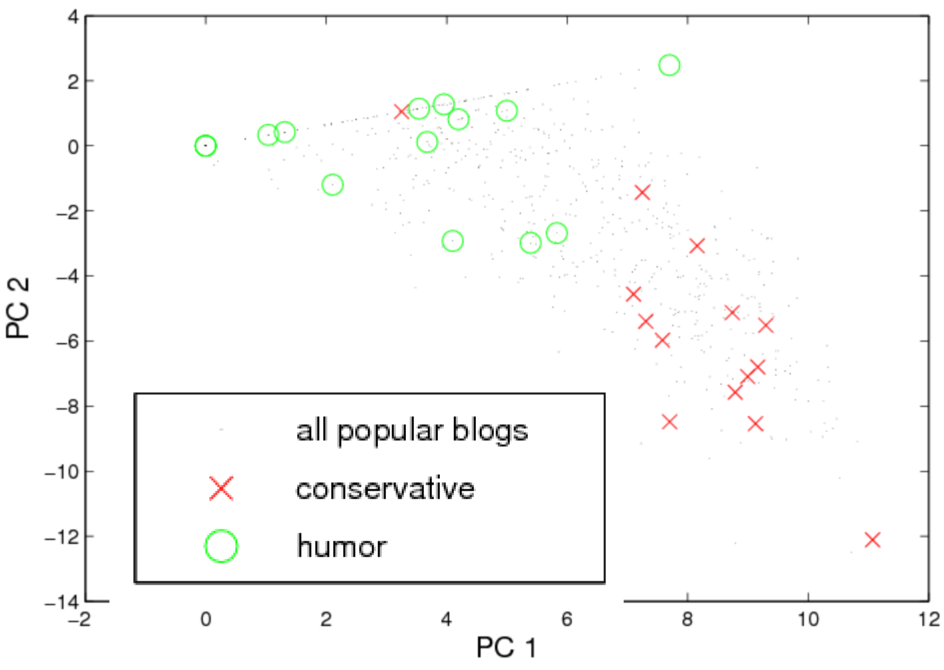
slashdot	4.6	2.1	.09			
boingboing	3.2	1.1		3.4	.07	
. . .	4.2					
. . .	5.1					
. . .	2.1		1.1			
. . .	.67			.07		
. . .	.01					

PCA on cascade types

- Observation: Content of blogs and cascade behavior are often related.

- Distinct clusters for "conservative" and "humorous" blogs (hand-labeling).

M. McGlohon, J. Leskovec, C. Faloutsos, M. Hurst, N. Glance. Finding Patterns in Blog Shapes and Blog Evolution. ICWSM 2007.

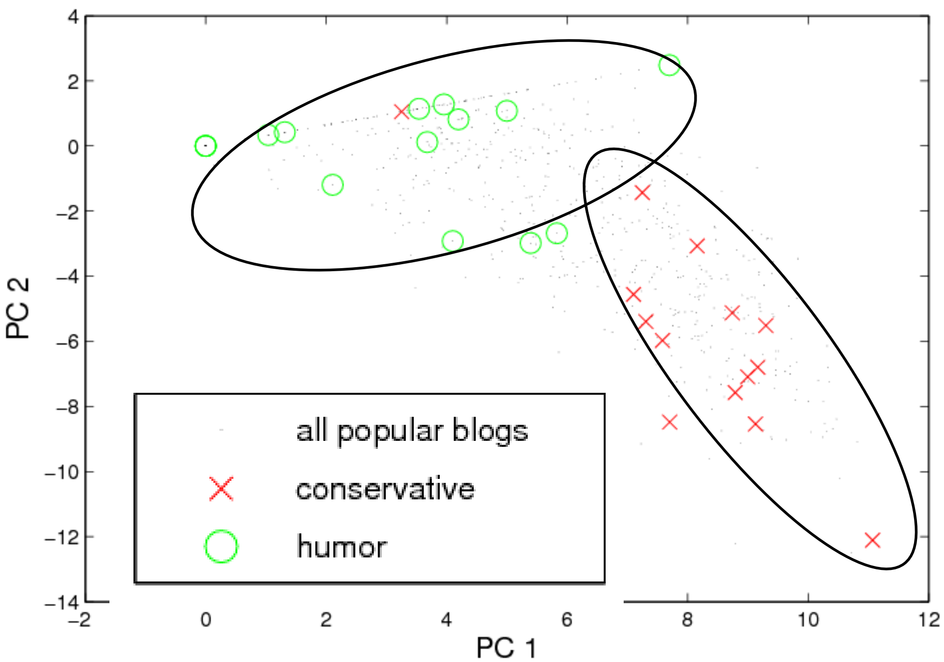


PCA on cascade types

- Observation: Content of blogs and cascade behavior are often related.

- Distinct clusters for "conservative" and "humorous" blogs (hand-labeling).

M. McGlohon, J. Leskovec, C. Faloutsos, M. Hurst, N. Glance. Finding Patterns in Blog Shapes and Blog Evolution. ICWSM 2007.



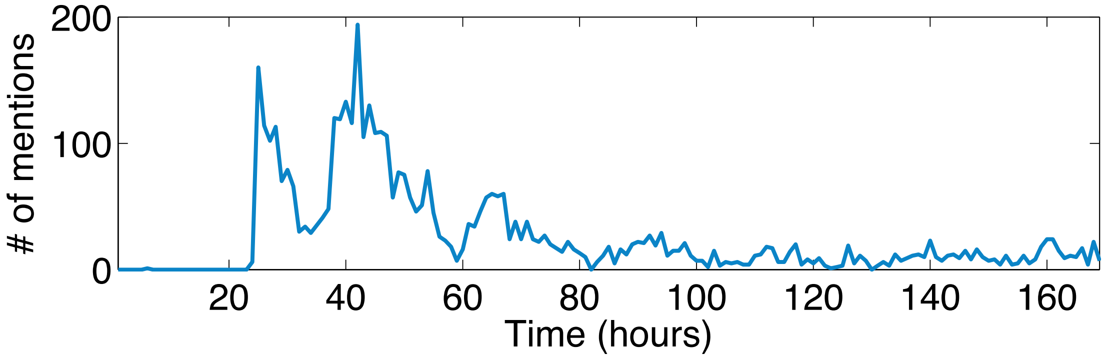
Part 3: Empirical Studies

- Q6: How do cascades look like?
- **Q7: How does activity evolve over time?**
- Q8: How does external influence act?

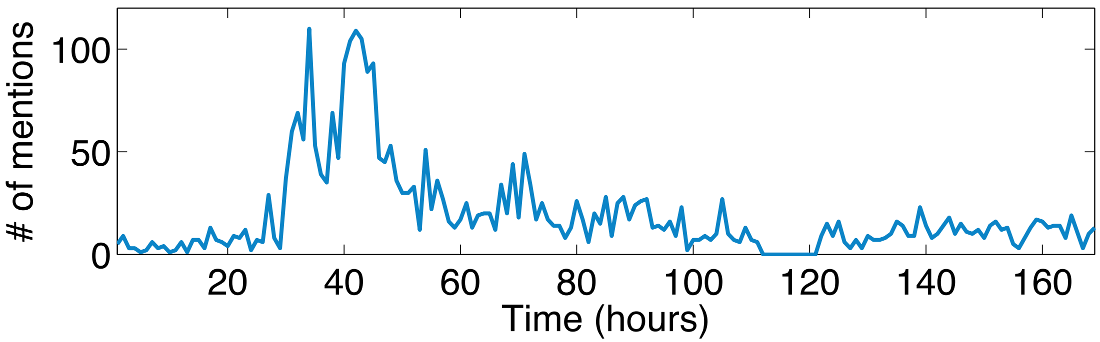
Rise and fall patterns in social media

- Meme (# of mentions in blogs)
  - short phrases Sourced from U.S. politics in 2008

"you can put lipstick on a pig"

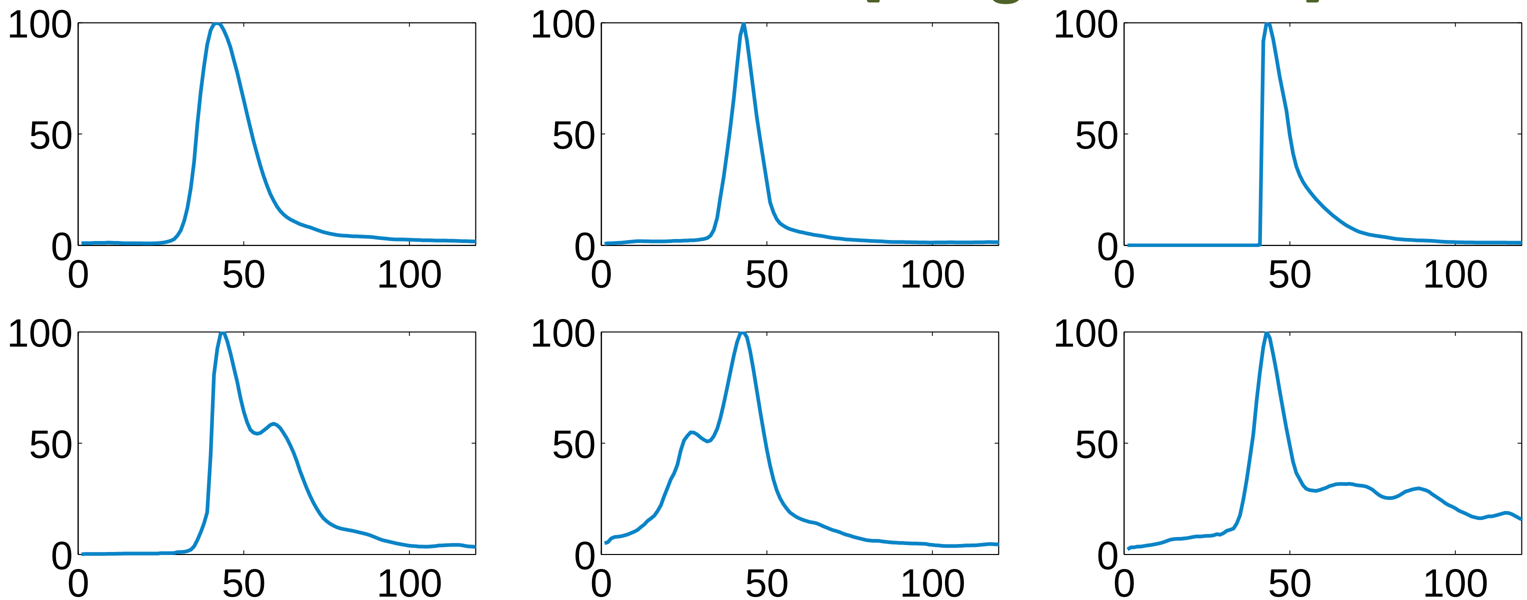


"yes we can"



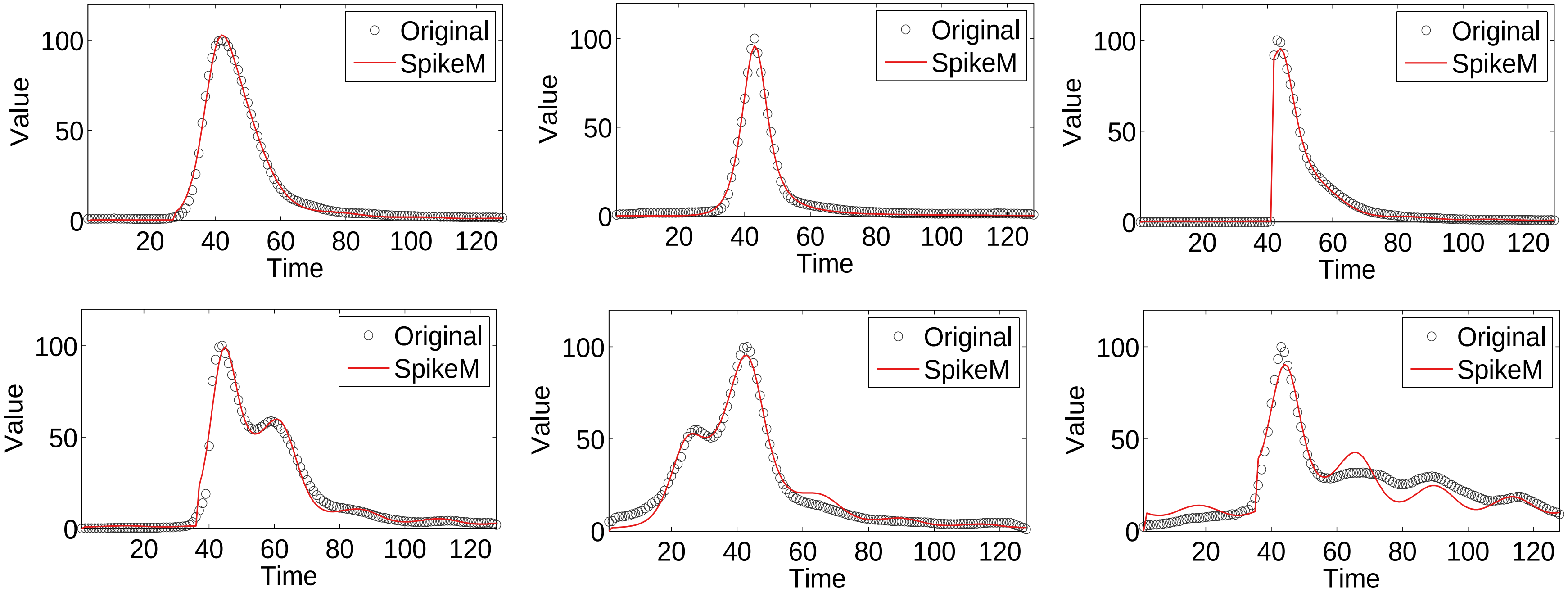
Rise and fall patterns in social media

- Can we find a unifying model, which includes these patterns?
    - **four** classes on YouTube [Crane et al. '08]
    - **six** classes on Meme [Yang et al. '11]



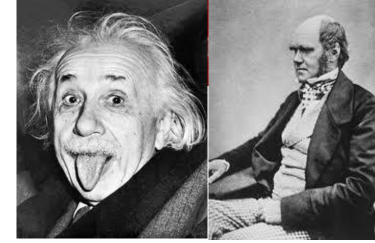
Rise and fall patterns in social media

- Answer: YES!



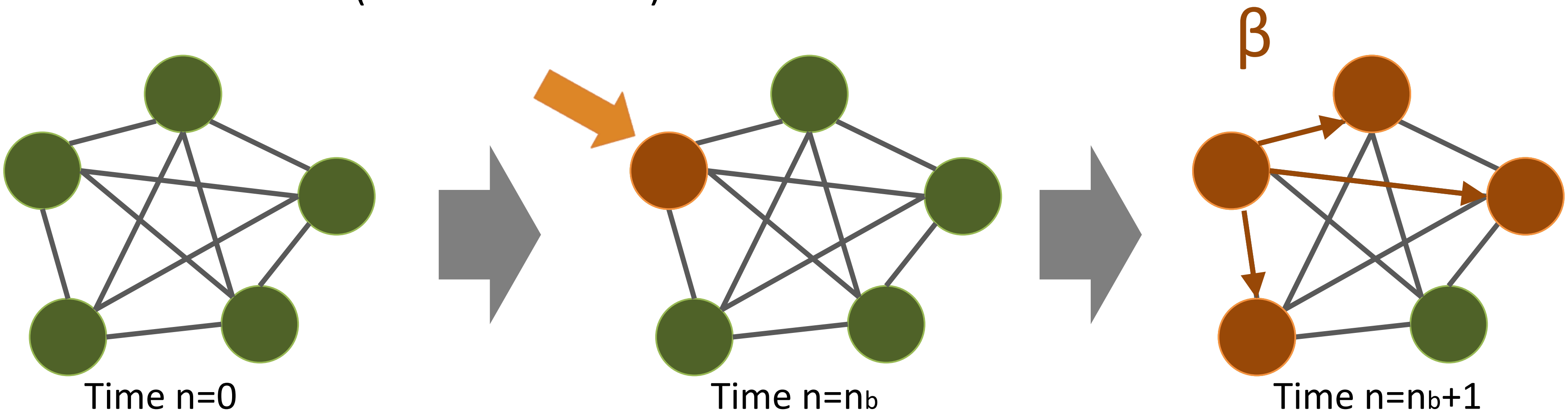
- We can represent all patterns by single model

In Matsubara+ SIGKDD 2012



Main idea - SpikeM

- 1. **Un**-informed bloggers (uninformed about rumor)
- 2. External shock at time n_b (e.g, breaking news)
- 3. Infection (word-of-mouth)



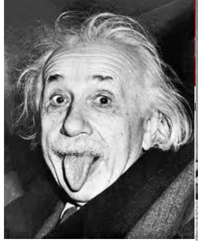
Infectiveness of a blog-post at age n: $f(n) = b * n^{-1.5}$

Power Law

b - Strength of infection (quality of news)

$f(n)$ - Decay function (how infective a blog posting is)

Prakash and Faloutsos 2012



-1.5 slope

J. G. Oliveira & A.-L. Barabási Human Dynamics: The Correspondence Patterns of Darwin and Einstein. *Nature* **437**, 1251 (2005) . [PDF]

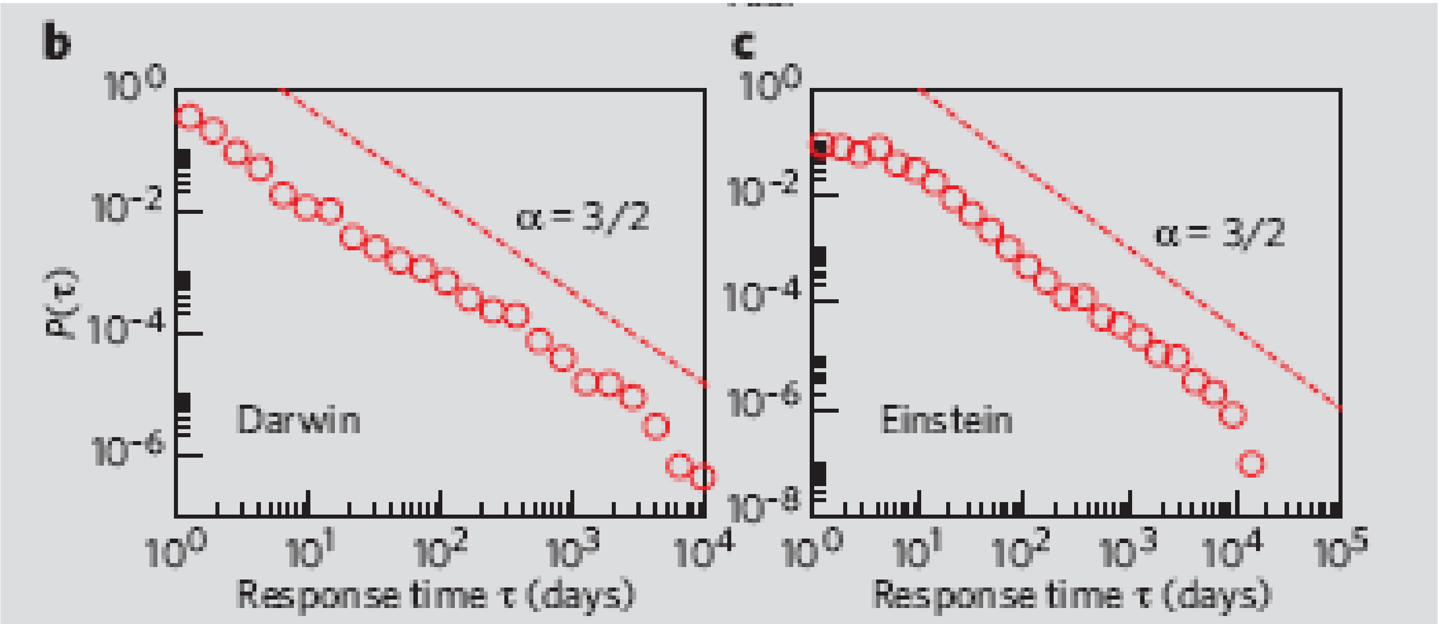


Figure 1 | The correspondence patterns of Darwin and Einstein.

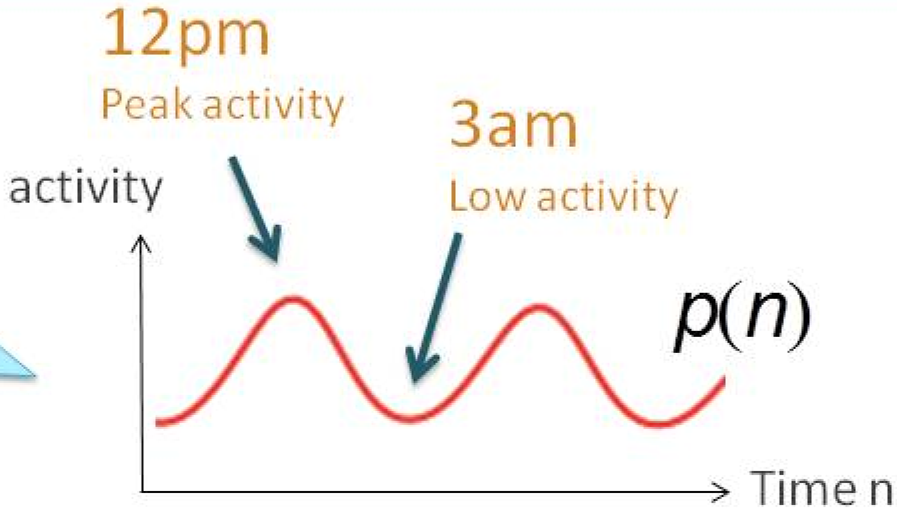
Details

SpikeM - with periodicity

- Full equation of SpikeM

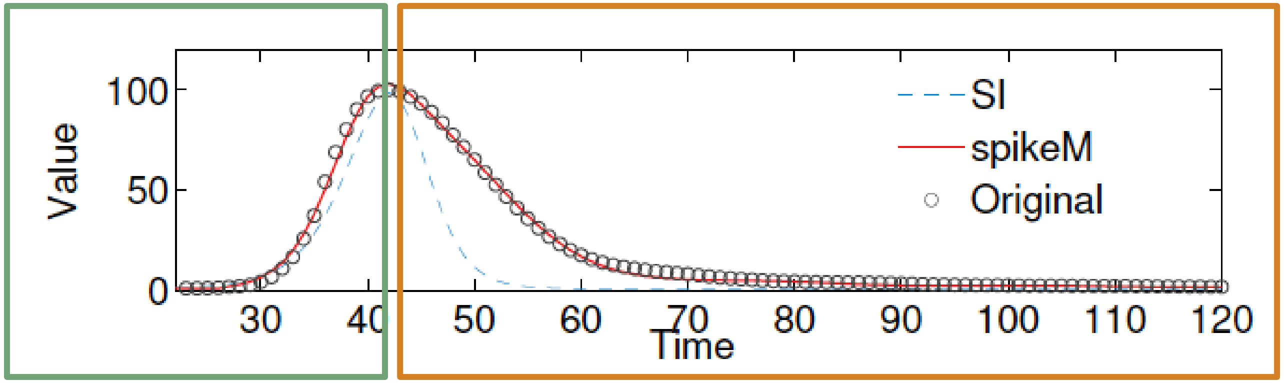
$$\Delta B(n+1) = \underbrace{p(n+1)}_{\text{Periodicity}} \cdot \left[U(n) \cdot \sum_{t=n_b}^{n} (\Delta B(t) + S(t)) \cdot f(n+1-t) + \varepsilon \right]$$

Bloggers change their activity over time
(e.g., daily, weekly, yearly)



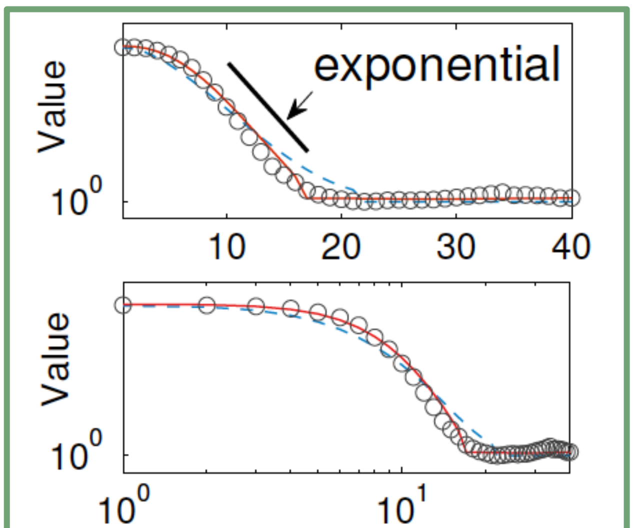
Details

- Analysis – exponential rise and power-raw fall



Liner-log

Log-log



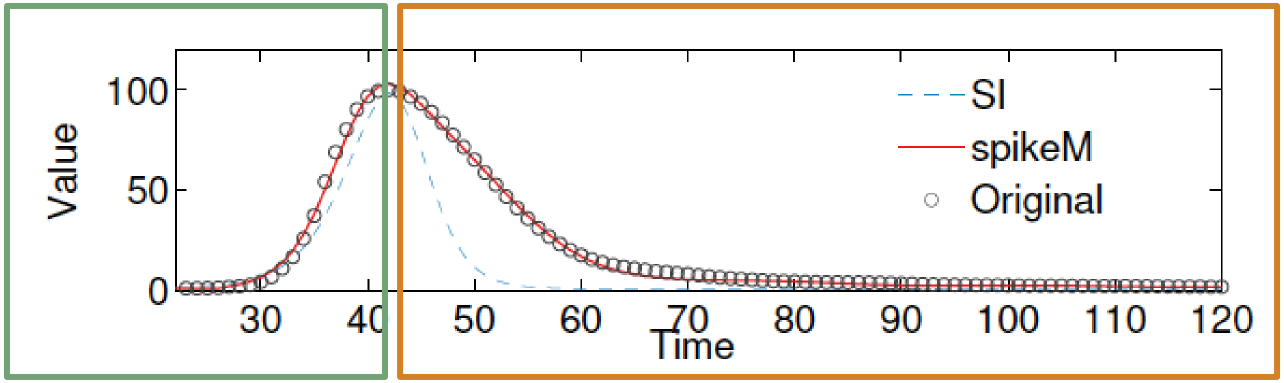
Rise-part

SI -> exponential

SpikeM -> exponential

Details

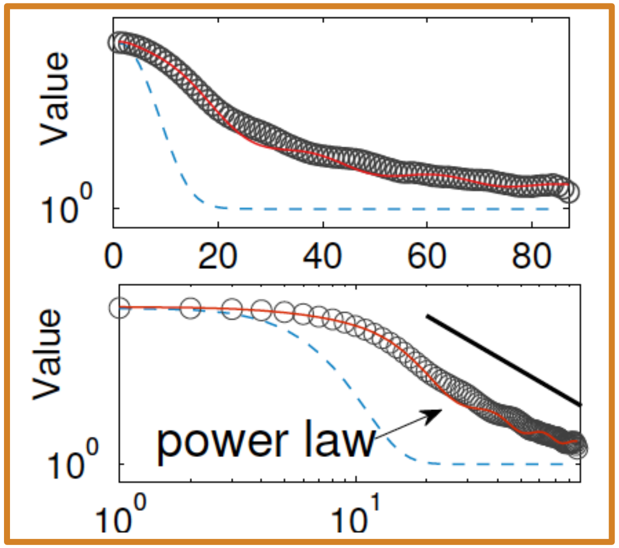
- Analysis – exponential rise and power-raw fall



Fall-part

❌ SI -> exponential

SpikeM -> power law

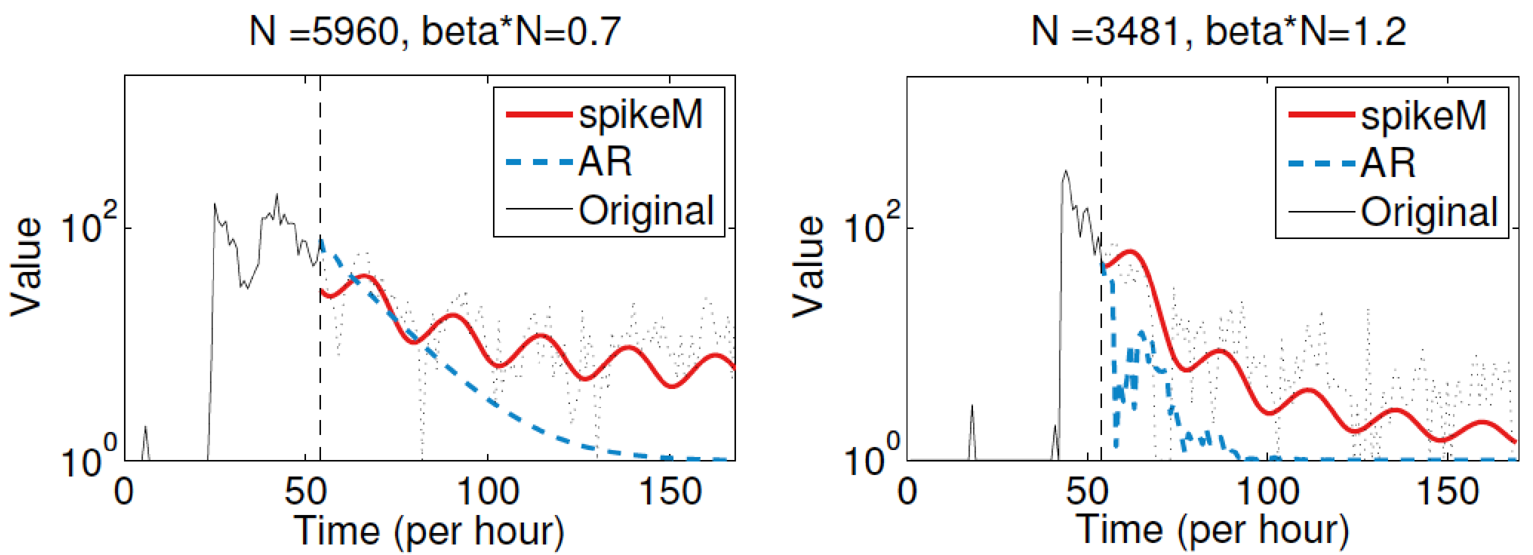


Liner-log

Log-log

Tail-part forecasts

- **SpikeM** can capture tail part

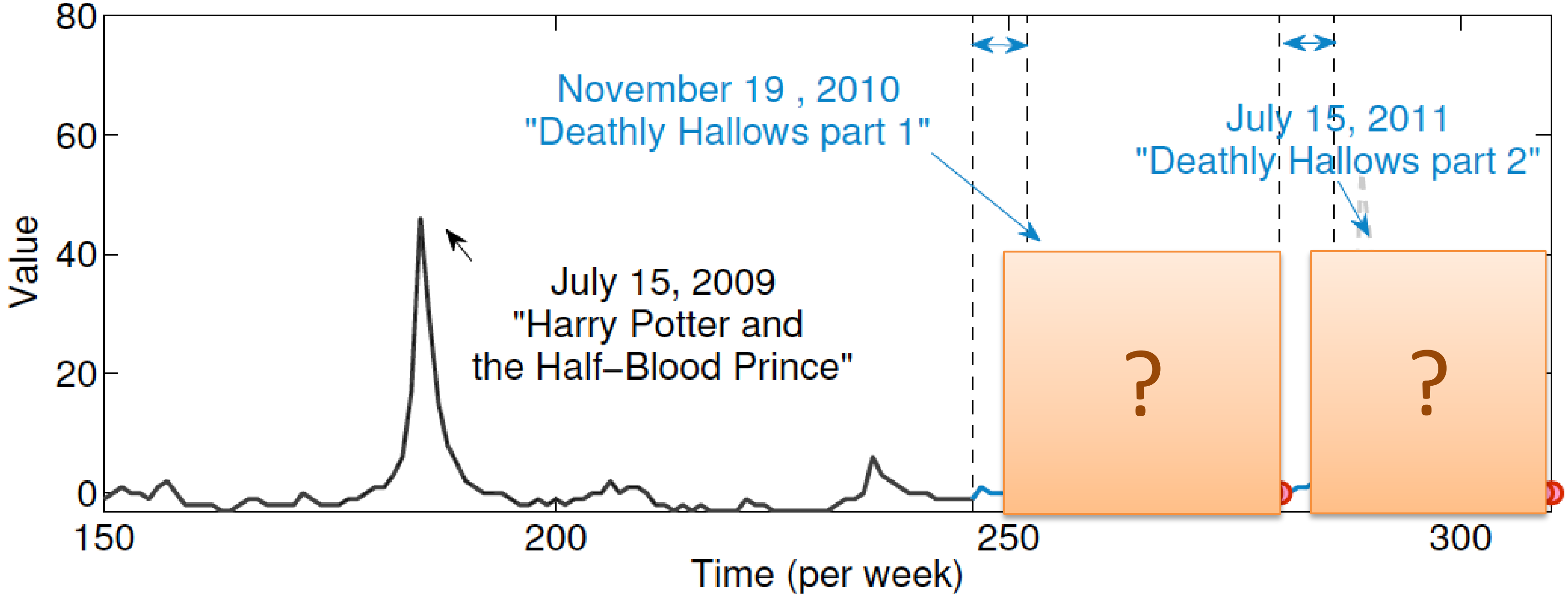


VirginiaTech

Carnegie Mellon

"What-if" forecasting

(1) First spike (2) Release date (3) Two weeks before release



e.g., given (1) first spike,

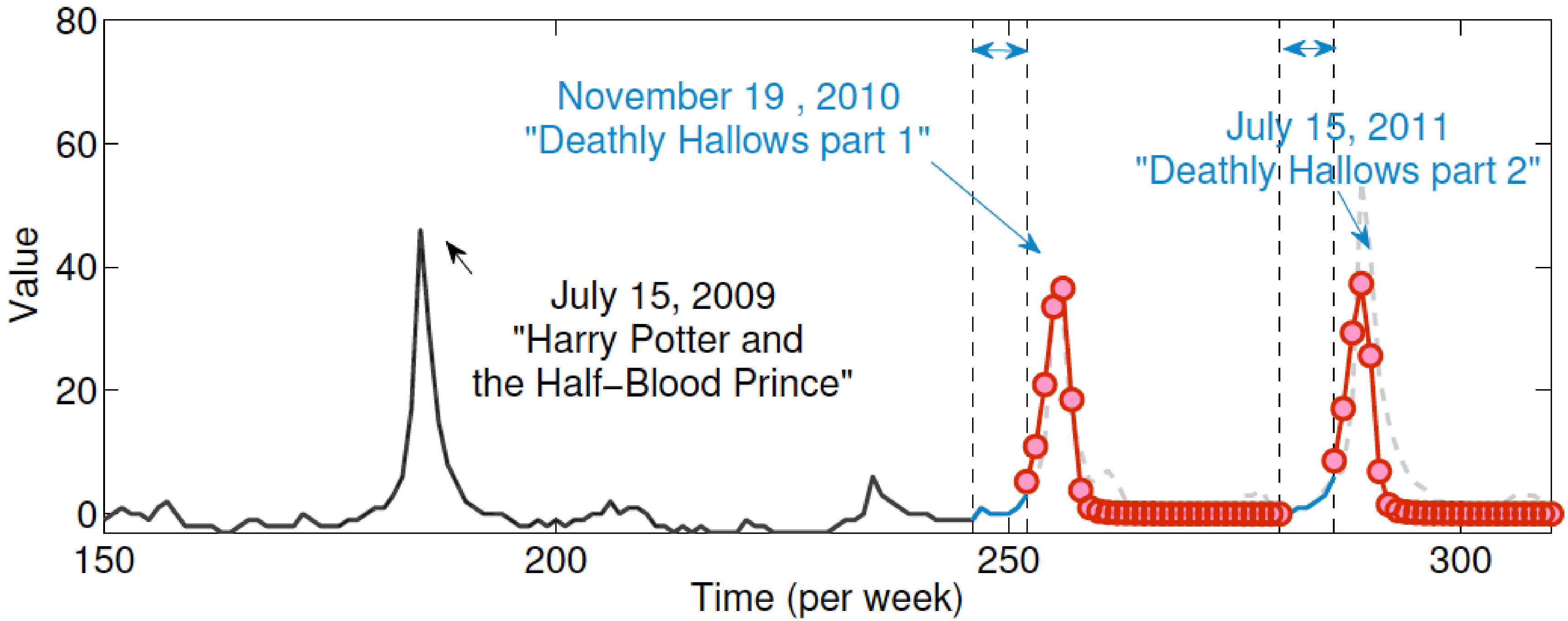
(2) release date of two sequel movies

(3) access volume before the release date

"What-if" forecasting

–SpikeM can forecast not only tail-part, but also **rise-part**!

(1) First spike (2) Release date (3) Two weeks before release



- **SpikeM** can forecast upcoming spikes

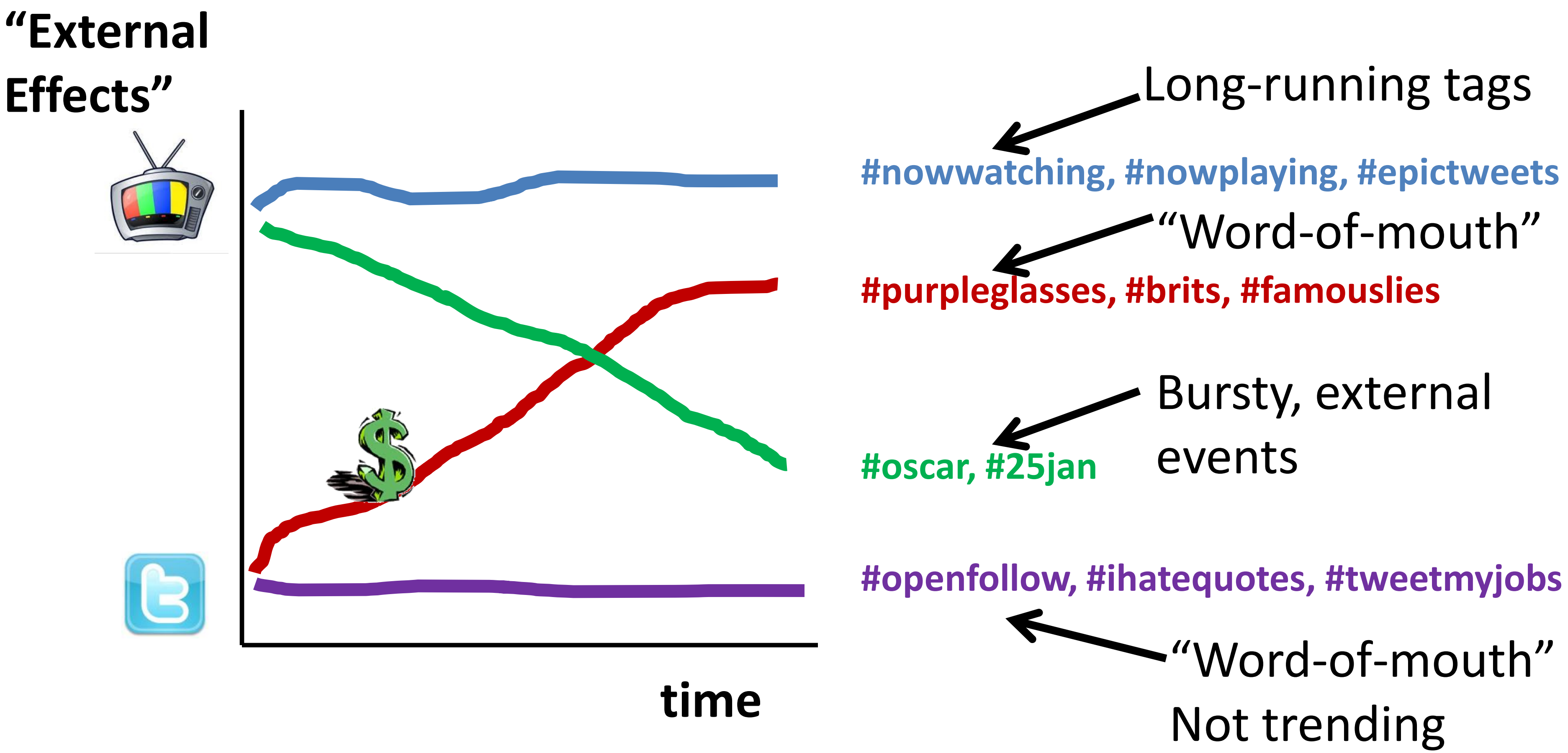
Part 3: Empirical Studies

- Q6: How do cascades look like?
- Q7: How does activity evolve over time?
- **Q8: How does external influence act?**

Tweets Diffusion: Problem Definition

- Given:
  - Action log of people tweeting a #hashtag ( )
  - A network of users ( )
- Find:
  - How external influence varies with #hashtags?

Results: External Influence vs Time



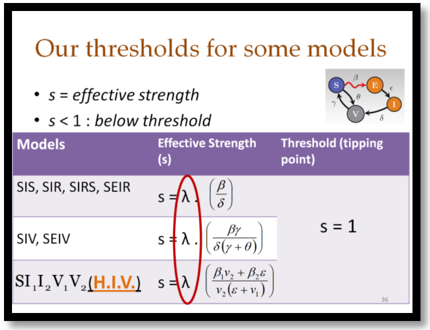
Can also use for Forecasting, Anomaly Detection!

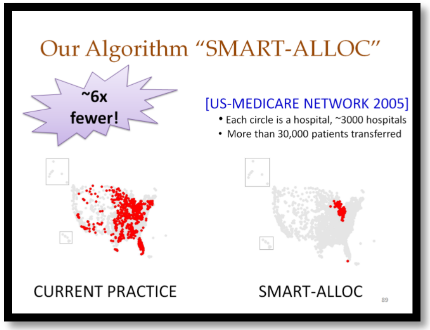
Outline

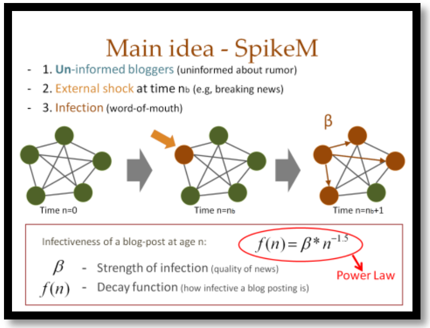
- Motivation
- Part 1: Understanding Epidemics (Theory)
- Part 2: Policy and Action (Algorithms)
- Part 3: Learning Models (Empirical Studies)
- **Conclusion**

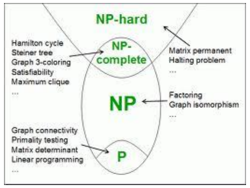
Conclusions

- Epidemic Threshold
  – It's the Eigenvalue
- Fast Immunization
  – Max. drop in eigenvalue, linear-time near-optimal algorithm
- Bursts: SpikeM model
  – Exponential growth, Power-law decay



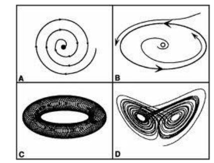


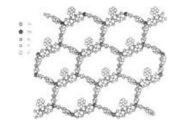




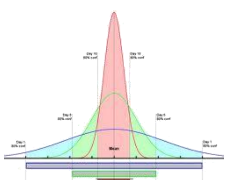
Propagation on Networks

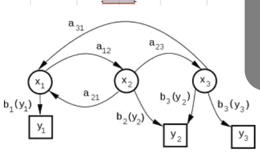
- Theory & Algo.
- Biology
- Physics
- Comp. Systems
- Social Science
- ML & Stats.
- Econ.

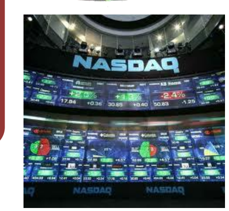




facebook







References

1. *Winner-takes-all: Competing Viruses or Ideas on fair-play networks (**B. Aditya Prakash**, Alex Beutel, Roni Rosenfeld, Christos Faloutsos)* – In WWW 2012, Lyon
2. *Threshold Conditions for Arbitrary Cascade Models on Arbitrary Networks (**B. Aditya Prakash**, Deepayan Chakrabarti, Michalis Faloutsos, Nicholas Valler, Christos Faloutsos)* - In IEEE ICDM 2011, Vancouver (*Invited to KAIS Journal **Best Papers of ICDM**.)*
3. *Times Series Clustering: Complex is Simpler! (Lei Li, **B. Aditya Prakash**)* - In ICML 2011, Bellevue
4. *Epidemic Spreading on Mobile Ad Hoc Networks: Determining the Tipping Point (Nicholas Valler, **B. Aditya Prakash**, Hanghang Tong, Michalis Faloutsos and Christos Faloutsos)* – In IEEE NETWORKING 2011, Valencia, Spain
5. *Formalizing the BGP stability problem: patterns and a chaotic model **(B. Aditya Prakash**, Michalis Faloutsos and Christos Faloutsos)* – In IEEE INFOCOM NetSciCom Workshop, 2011.
6. *On the Vulnerability of Large Graphs (Hanghang Tong, **B. Aditya Prakash**, Tina Eliassi-Rad and Christos Faloutsos)* – In IEEE ICDM 2010, Sydney, Australia
7. *Virus Propagation on Time-Varying Networks: Theory and Immunization Algorithms (**B. Aditya Prakash**, Hanghang Tong, Nicholas Valler, Michalis Faloutsos and Christos Faloutsos)* – In ECML-PKDD 2010, Barcelona, Spain
8. *MetricForensics: A Multi-Level Approach for Mining Volatile Graphs (Keith Henderson, Tina Eliassi-Rad, Christos Faloutsos, Leman Akoglu, Lei Li, Koji Maruhashi**, B. Aditya Prakash** and Hanghang Tong*) - In SIGKDD 2010, Washington D.C.
9. *Parsimonious Linear Fingerprinting for Time Series* (*Lei Li, **B. Aditya Prakash** and Christos Faloutsos*) - In VLDB 2010, Singapore
10. *EigenSpokes: Surprising Patterns and Scalable Community Chipping in Large Graphs (**B. Aditya Prakash**, Ashwin Sridharan, Mukund Seshadri, Sridhar Machiraju and Christos Faloutsos)* – In PAKDD 2010, Hyderabad, India
11. *BGP-lens: Patterns and Anomalies in Internet-Routing Updates (**B. Aditya Prakash**, Nicholas Valler, David Andersen, Michalis Faloutsos and Christos Faloutsos)* – In ACM SIGKDD 2009, Paris, France.
12. *Surprising Patterns and Scalable Community Detection in Large Graphs (**B. Aditya Prakash**, Ashwin Sridharan, Mukund Seshadri, Sridhar Machiraju and Christos Faloutsos)* – In IEEE ICDM Large Data Workshop 2009, Miami
13. *FRAPP: A Framework for high-Accuracy Privacy-Preserving Mining (Shipra Agarwal, Jayant R. Haritsa and **B. Aditya Prakash**)* – In Intl. Journal on Data Mining and Knowledge Discovery (DKMD), Springer, vol. 18, no. 1, February 2009, Ed: Johannes Gehrke.
14. *Complex Group-By Queries For XML (C. Gokhale, N. Gupta, P. Kumar, L. V. S. Lakshmanan, R. Ng and **B. Aditya Prakash**)* – In IEEE ICDE 2007, Istanbul, Turkey.

Acknowledgements

Collaborators

Christos Faloutsos
Roni Rosenfeld,
Michalis Faloutsos, UC RIVERSIDE
Lada Adamic, MICHIGAN
Theodore Iwashyna (M.D.), MICHIGAN
Dave Andersen,
Tina Eliassi-Rad, RUTGERS
Iulian Neamtiu, UC RIVERSIDE
Varun Gupta, CHICAGO BOOTH
Jilles Vreeken, Universiteit Antwerpen

Deepayan Chakrabarti, YAHOO!
Hanghang Tong, IBM
Kunal Punera, YAHOO!
Ashwin Sridharan, Sprint
Sridhar Machiraju, Google
Mukund Seshadri, amazon
Alice Zheng, Microsoft
Lei Li, Cal
Polo Chau,
Nicholas Valler, UC RIVERSIDE
Alex Beutel,
Xuetao Wei UC RIVERSIDE

Acknowledgements

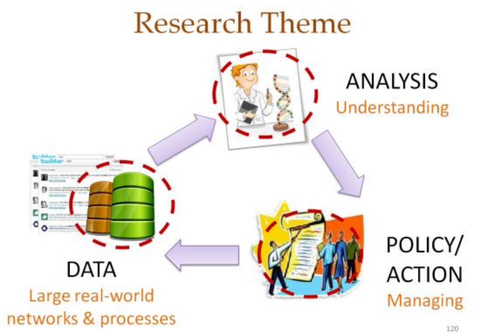
Funding

INARC
DARPA
NSF
YAHOO!
IBM
Sprint
Microsoft

Dynamical Processes on Large Networks

B. Aditya Prakash

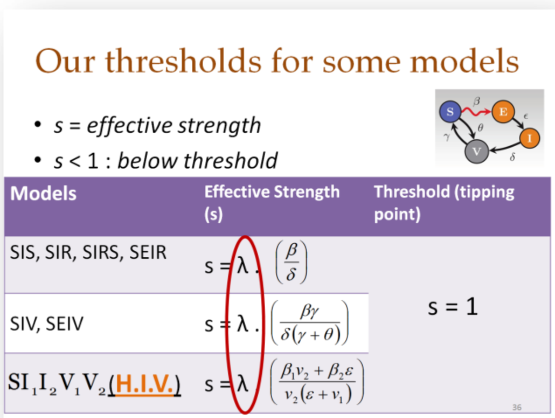
Christos Faloutsos



Analysis

Policy/Action

Data



Our thresholds for some models

- s = effective strength
- s < 1 : below threshold

Models	Effective Strength (s)	Threshold (tipping point)
SIS, SIR, SIRS, SEIR	$s = \lambda \cdot \left(\frac{\beta}{\delta}\right)$	s = 1
SIV, SEIV	$s = \lambda \cdot \left(\frac{\beta\gamma}{\delta(\gamma+\theta)}\right)$	
$SI_1I_2V_1V_2$ (H.I.V.)	$s = \lambda \cdot \left(\frac{\beta_1 v_2 + \beta_2 \varepsilon}{v_2(\varepsilon + v_1)}\right)$