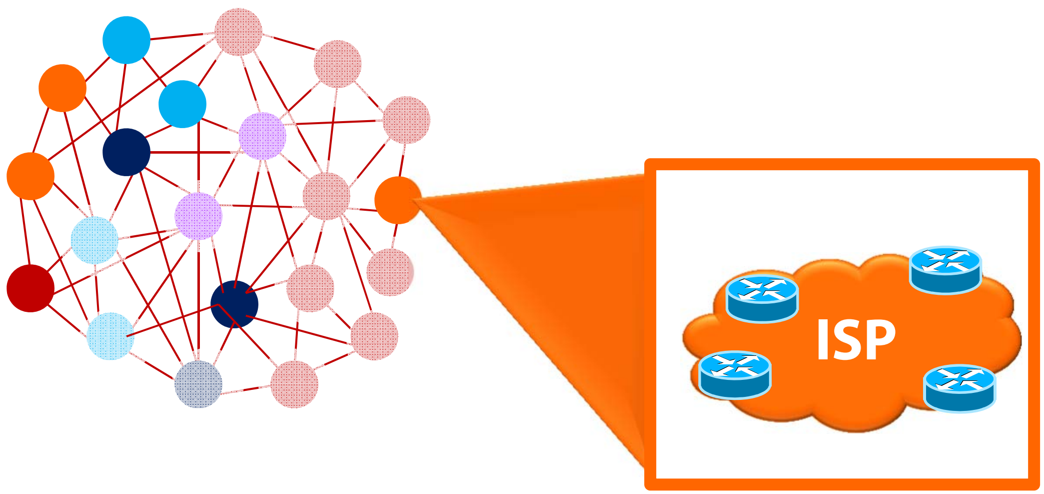
Diffusion of Networking Technologies



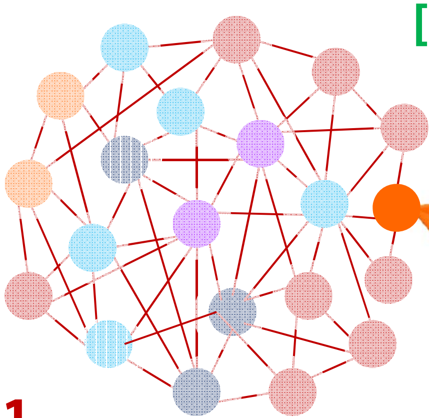
Bellairs Workshop on Algorithmic Game Theory
Barbados April 2012

Sharon Goldberg
Boston University

Zhenming Liu
Harvard University

Diffusion in social networks: Linear Threshold Model

[Kempe Kleinberg Tardos'03, Morris'01, Granovetter'78]



A node's utility depends only on its neighbors!

I'll adopt the innovation if θ of my friends do!

$\theta = 1$
$\theta = 2$
$\theta = 3$
$\theta = 4$
$\theta = 6$

Optimization problem [KKT'03]: Given the graph and thresholds, what is the smallest seedset that can cause the entire network to adopt?

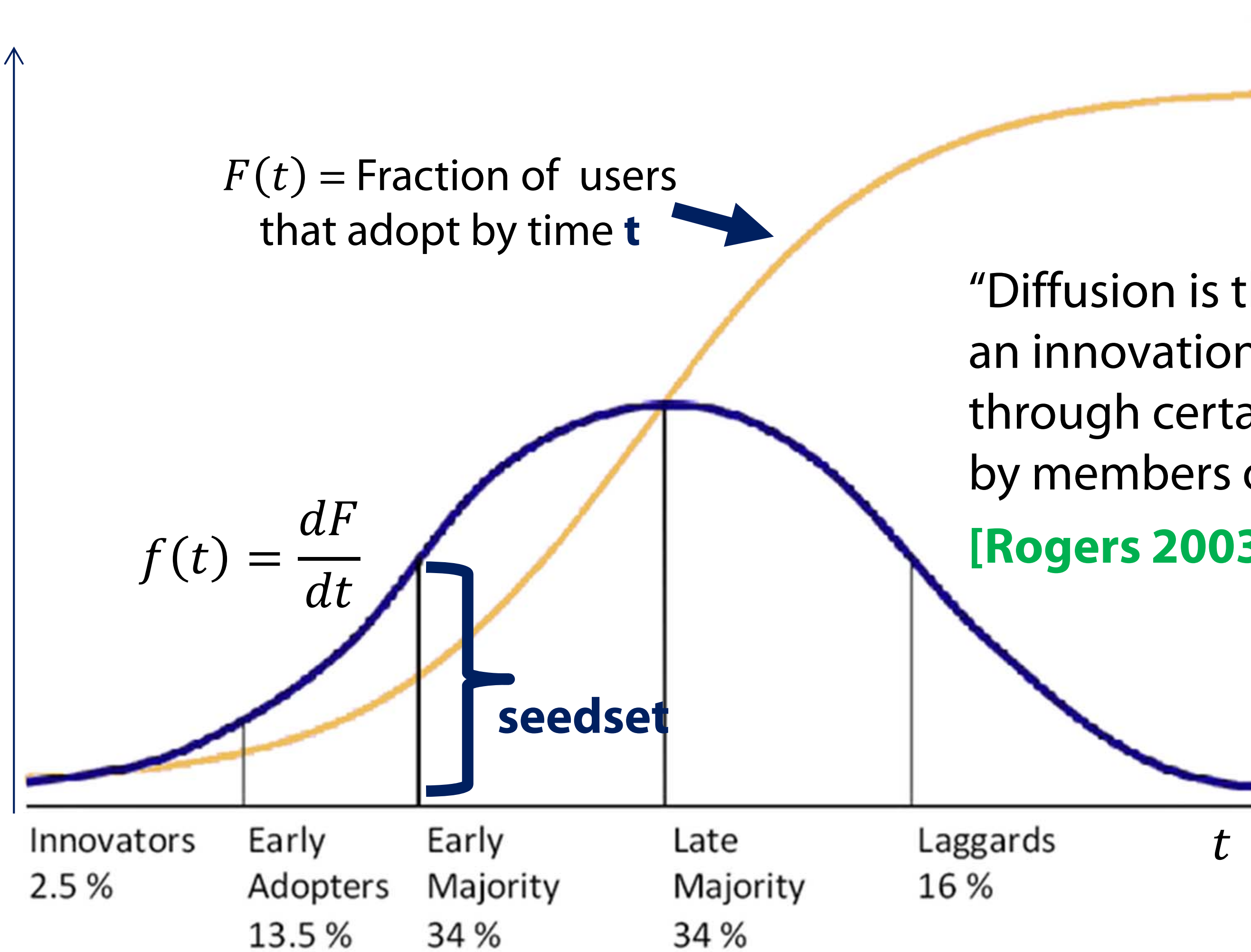
Seedset: A set of nodes that can kick off the process.
Marketers, policy makers, and spammers can target them as early adopters!

What if the **innovation** is a **networking technology** (e.g. IPv6, Secure BGP, QoS, etc)

And the **graph** is the network?

Inspiration: The literature on diffusion of innovations (1)

- **Social Sciences:** **[Ryan and Gross'49, Rogers '62, ….]**
  - General theory tested empirically in different settings (corn, Internet, etc)



"Diffusion is the process by which an innovation is communicated through certain channels over time by members of a social system."

[Rogers 2003]

Image: Wikipedia

Inspiration: The literature on diffusion of innovations (2)

- **Social Sciences:** **[Ryan and Gross'49, Rogers '62, ….]**
  - General theory tested empirically in different settings (corn, Internet, etc)

- **Marketing:** The Bass Model **[Bass'69]**
  - Forecasting extent of diffusion, and how pricing, marketing mix effects it

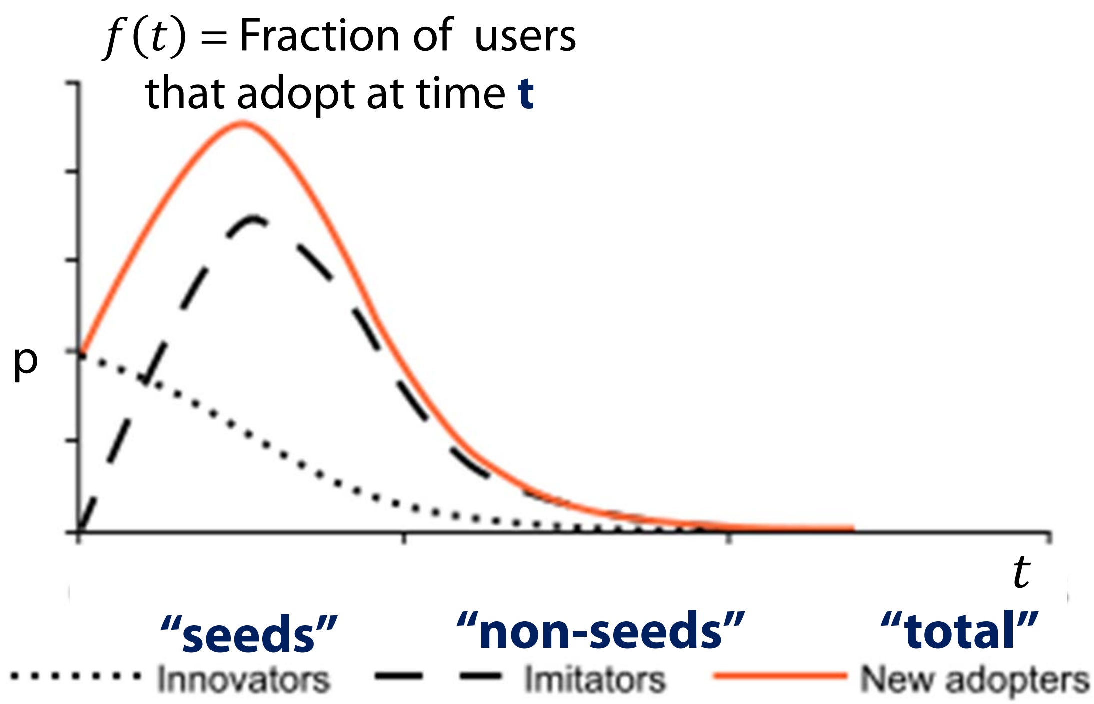


Image: Wikipedia

Inspiration: The literature on diffusion of innovations (3)

- **Social Sciences: [Ryan and Gross'49, Rogers '62, ….]**
  - General theory tested empirically in different settings (corn, Internet, etc)
- **Marketing:** The Bass Model **[Bass'69]**
  - Forecasting extent of diffusion, and how pricing, marketing mix effects it
- **Economics:** "Network externalities" or "Network effects" **[Katz Shapiro'85…]**
  - Models to analyze markets, econometric validation, etc

"The utility that a given user derives from the good depends upon the **number** of other users who are in the same "network" as he or she."

[Katz & Shapiro 1985]

Inspiration: The literature on diffusion of innovations (4)

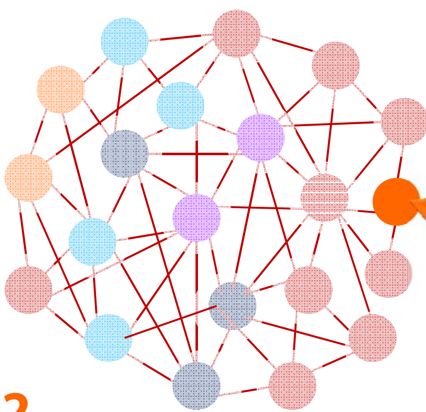
- **Social Sciences:** **[Ryan and Gross'49, Rogers '62, ...]**
  - General theory tested empirically in different settings (corn, Internet, etc)
- **Marketing:** The Bass Model **[Bass'69]**
  - Forecasting extent of diffusion, and how pricing, marketing mix effects it
- **Economics:** "Network externalities" or "Network effects" **[Katz Shapiro'85…]**
  - Models to analyze markets, econometric validation, etc
- **Popular Science:** "Metcalfe's Law" **[Metcalfe 1995]**

Traditional work: No graph. Utility depends on number of adopters.

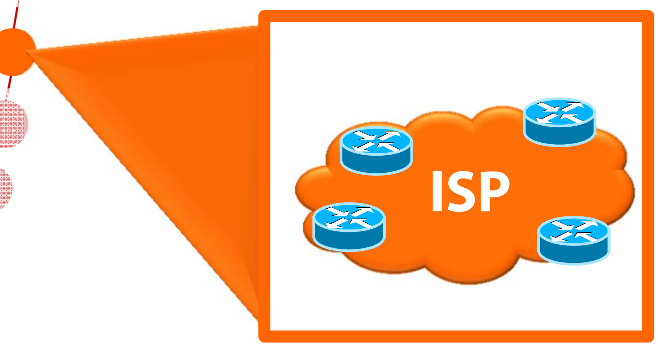
[KKT'03, …]: The graph is a social network. Utility is **local**.

Our model: Graph is an internetwork. Utility is **non-local**.

Diffusion in Internetworks: A new, non-local model (1)



Network researchers have been trying to understand why its so hard to deploy new technologies (**IPv6**, **secure BGP,** etc.)

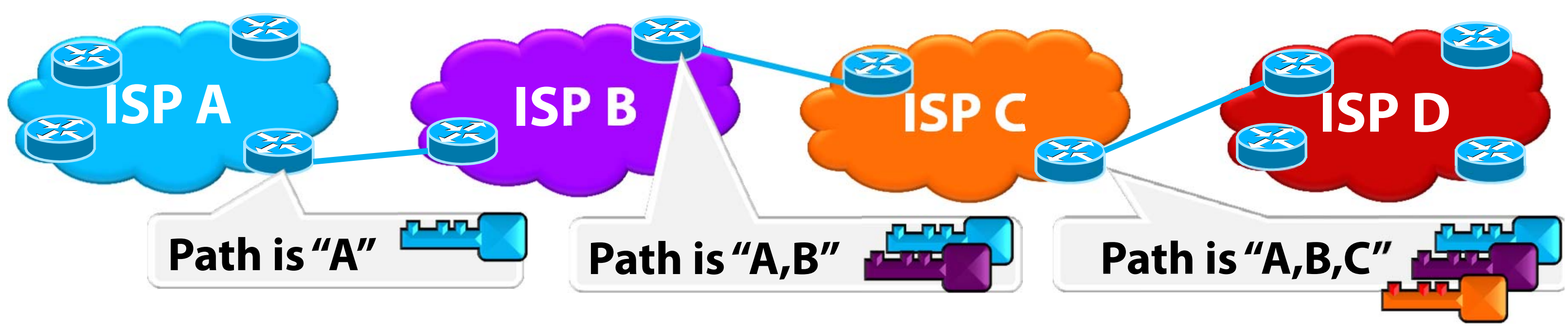


I'll adopt the innovation if I can use it to communicate with at least θ other Internet Service Providers (ISPs)!

θ = 2
θ = 3
θ = 12
θ = 15
θ = 16

These technologies work only if **all nodes on a path** adopt them.

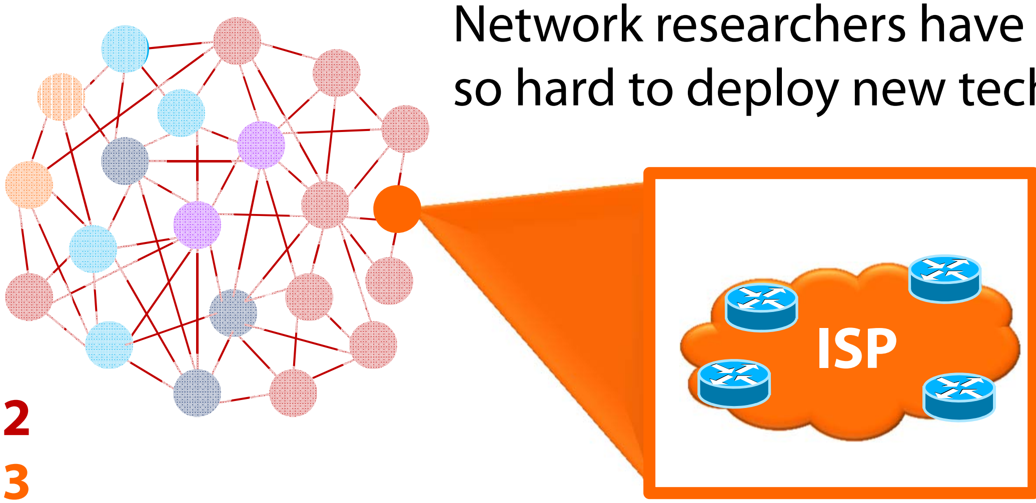
e.g. **Secure BGP** (Currently being standardized.)
All nodes must cryptographically sign messages so path is secure.



Other technologies share this property: QoS, fault localization, IPv6, …

Diffusion in internetworks: A new, non-local model (2)

Network researchers have been trying to understand why its so hard to deploy new technologies (**IPv6**, **secure BGP**, etc.)

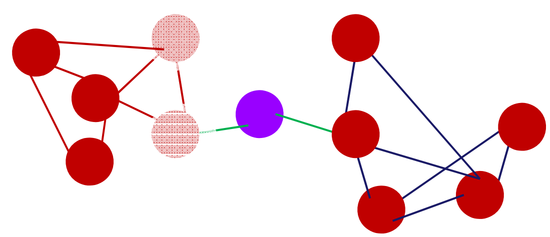


I'll adopt the innovation if I can use it to communicate with at least θ other Internet Service Providers (ISPs)!

θ = 2
θ = 3
θ = 12
θ = 15
θ = 16

Our new model of node utility: Node **u**'s utility depends on the size of the connected component of active nodes that **u** is part of.

eg. utility(u) = 5



Seedset: A set of nodes that can kick off the process.
Policy makers, regulatory groups can target them as early adopters!

Optimization problem: Given the graph and thresholds,
what is the smallest seedset that can cause the entire network to adopt?

Social networks (Local) vs Internetworks (Non-Local)

Minimization formulation: Given the graph and thresholds **θ**, find the smallest seedset that activates every node in the graph.

Local influence: **Deadly hard!**

Thm [Chen'08]: Finding an $O(2^{\log^{1-\varepsilon}|V|})$-approximation is NP hard.

Non-Local influence (Our model!): **Much less hard.**

Our main result: An $O(r \cdot k \cdot \log |V|)$ approx algorithm

---

Maximization formulation: Given the graph, assume **θ's are drawn uniformly at random**. Find seedset of size **k** maximizing number of active nodes.

Local influence: **Easy!**

Thm [KKT'03]: An $O(1-1/e)$-approximation algorithm.

How? 1) Prove submodularity. 2) Apply greedy algorithm.

Non-Local influence (Our model!): **The usual submodularity tricks fail.**

Our Results

Minimization formulation: Given the graph and thresholds **θ**, find the smallest seedset that activates every node in the graph.

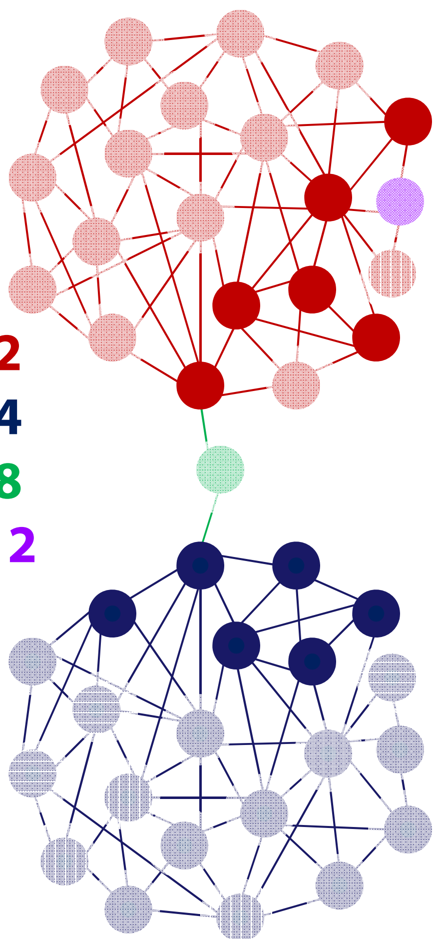
Main result: An $\mathbf{O(r \cdot k \cdot \log |V|)}$ approx algorithm

r is graph diameter (length of longest shortest path)
k is threshold granularity (number of thresholds)

Lower Bound: Can't do better than an $\mathbf{\Omega(\log |V|)}$ approx.
(Even for constant **r** and **k**.)

Lower Bound: Can't do better that an $\mathbf{\Omega(r)}$ approx. with our approach.

Terminology & Overview



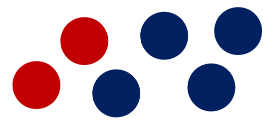
$\theta = 2$

$\theta = 4$

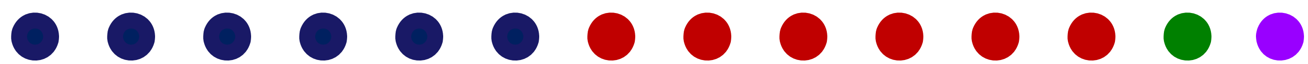
$\theta = 8$

$\theta = 12$

The problem: Given the graph and thresholds θ, find the smallest seedset that activates every node in the graph.

Seedset: 

Activation sequence:
(Time at which nodes activate, one per step)



Talk plan:

Part I: From global to local constraints

- Using connectivity.

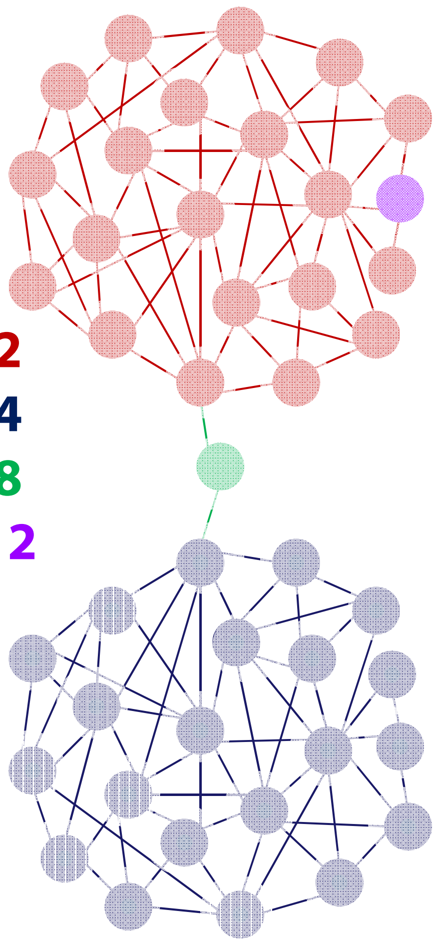
Part II: Approximation algorithm

Part I: From global to local.

(via a 2-approximation)

Why connectivity makes life better.

$\theta = 2$
$\theta = 4$
$\theta = 8$
$\theta = 12$



The trouble with disjoint components:
Activation of a distant node can dramatically change utility

utility(u) = 7 $\xrightarrow{\text{v activates}}$ **utility(u)= 15**

It's difficult to encode this with local constraints.

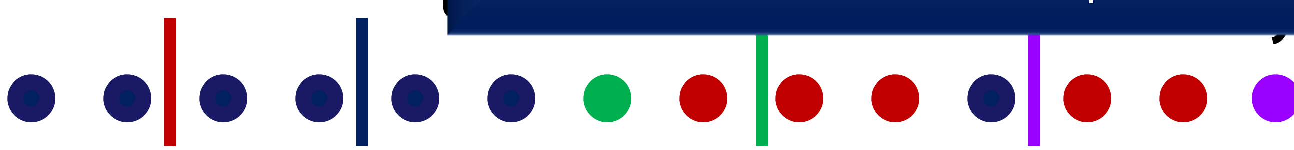
What if we search for connected activation sequences?
(There is a single connected active component at all times)

- Utility at activation = position in sequence
- To extract smallest seedset consistent with sequence:

Just check if $t > \theta$!

Thm: There is a **connected** activation sequence which has |seedset| < **2opt**.

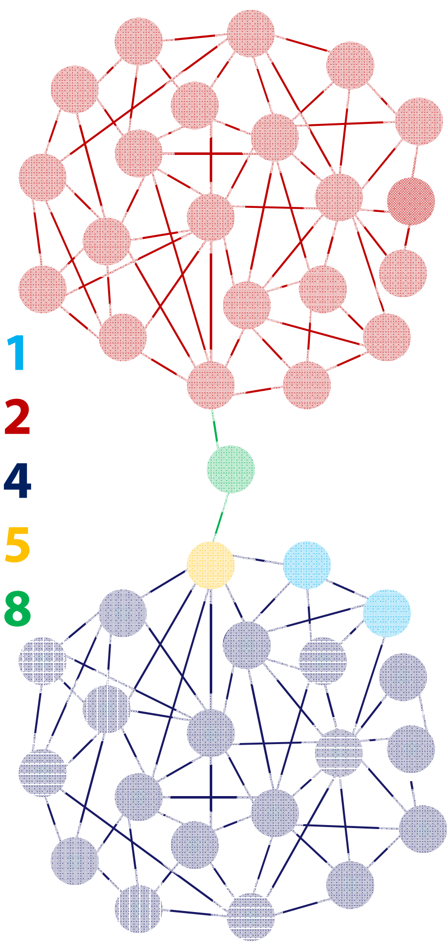
Activation sequence



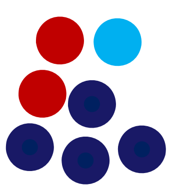
θ **θ ⇨ v is a seed** **θ** **θ ⇨ u is not a seed!**

Proof: ∃ connected sequence with |seedset| < 2opt. (1)

$\theta = 1$
$\theta = 2$
$\theta = 4$
$\theta = 5$
$\theta = 8$

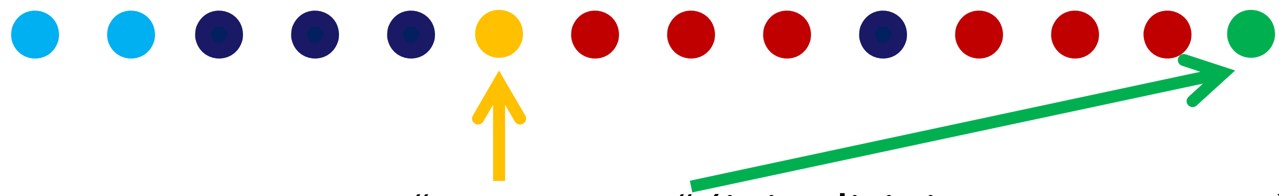


Seedset:



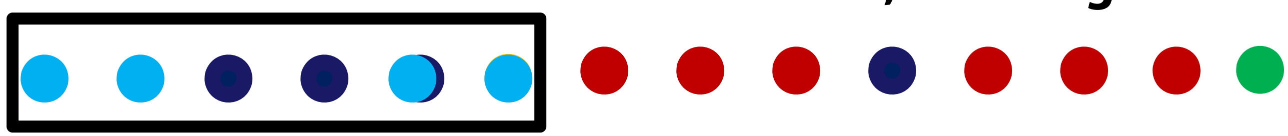
Proof: Given any **optimal sequence**
transform it to a **connected sequence**
by adding at most **opt** nodes to the seedset.

Optimal (disconnected) activation sequence



"connectors" (join disjoint components)

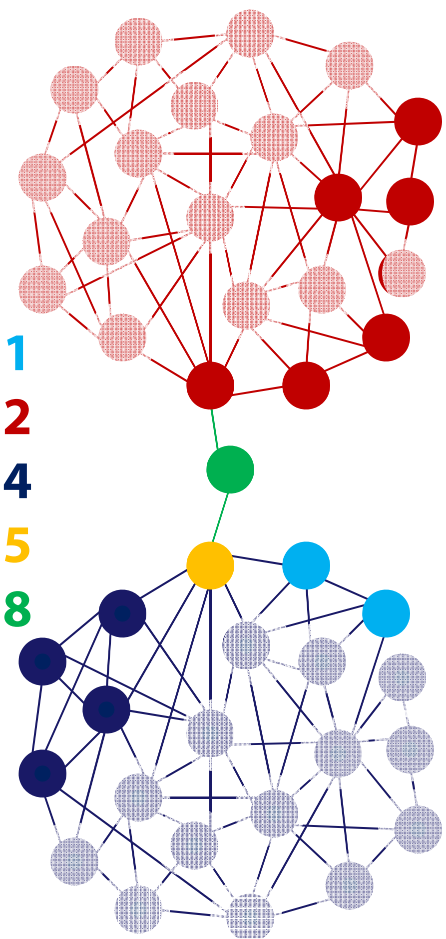
Transform: Add connector to seedset, rearrange



We always activate **large component** first.

Why? Non-seeds in **small component** must have θ smaller than size of **large component**
⇨ no non-connectors are added to seedset!

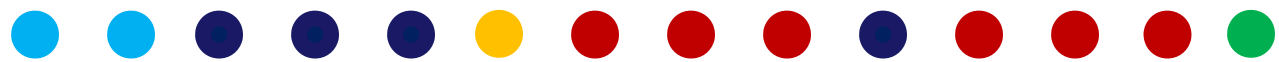
Proof: ∃ connected sequence with |seedset| < 2opt. (2)



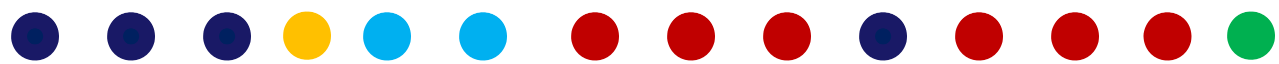
$\theta = 1$
$\theta = 2$
$\theta = 4$
$\theta = 5$
$\theta = 8$

Proof: Given any **optimal sequence**
transform it to a **connected sequence**
by adding at most **opt** nodes to the seedset.

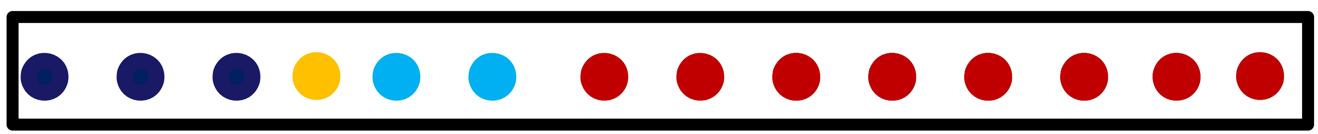
Optimal (disconnected) activation sequence



Transform: Add connector to seedset, rearrange

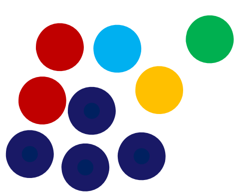


Transform: Add connector to seedset, rearrange

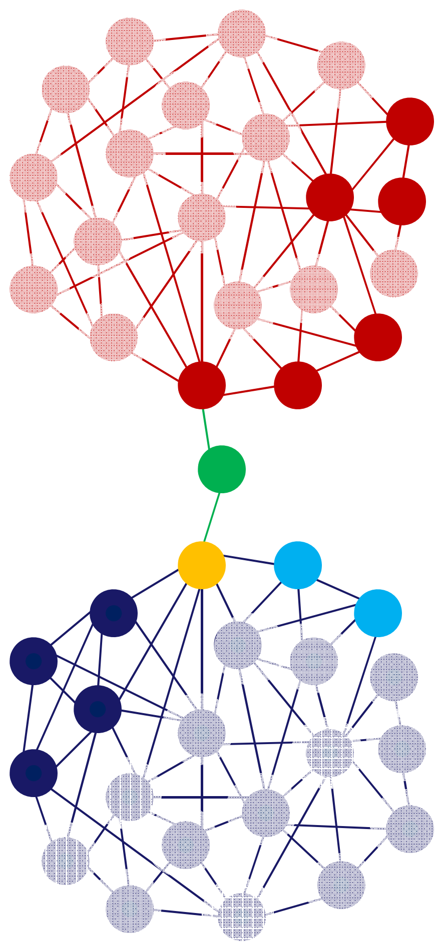


The activation sequence is now connected.

Seedset:

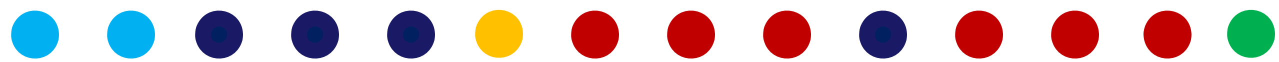


Proof: ∃ connected sequence with |seedset| < 2opt. (3)



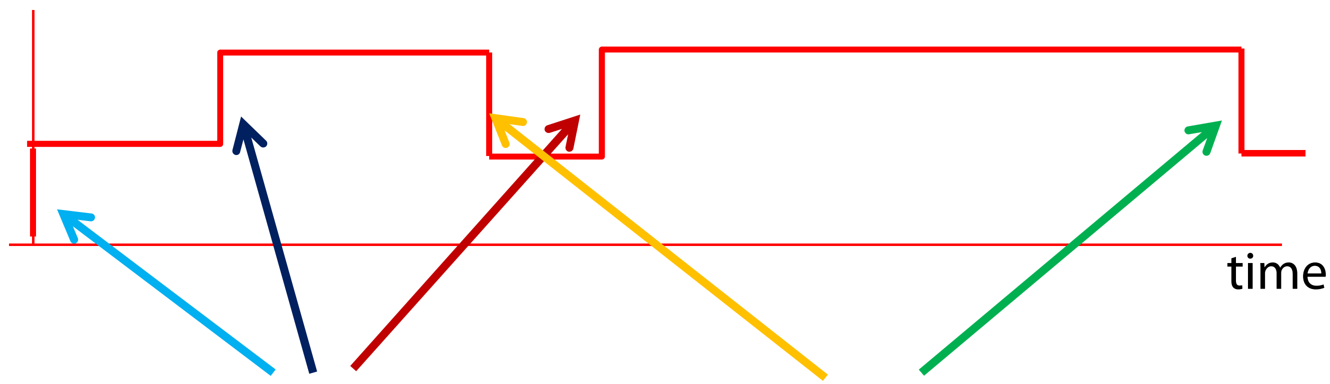
Proof: Given any **optimal sequence** transform it to a **connected sequence** by adding at most **opt** nodes to the seedset.

Optimal (disconnected) activation sequence



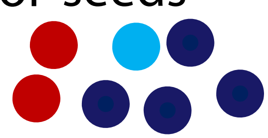
To bound seedset growth, we bound # of connectors.

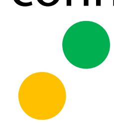
Plot of # of disconnected components in optimal sequence



Every step up needs a step down

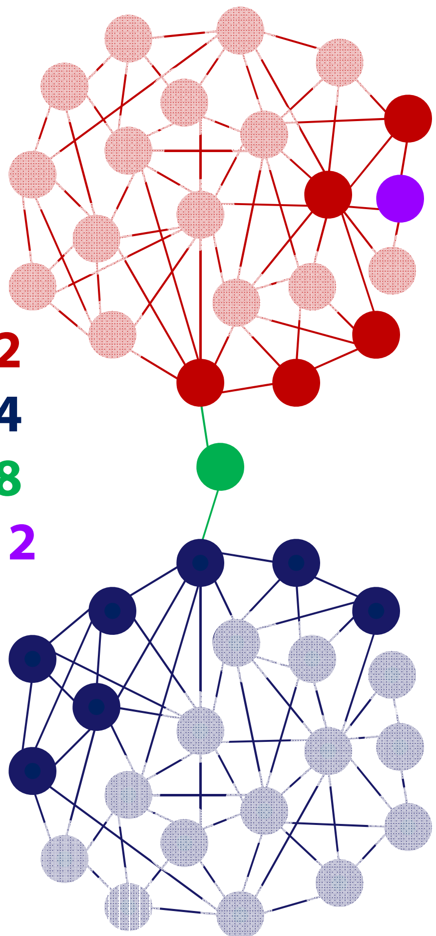
⇨ # of seeds > # of connectors





In the worst case, our transformation doubles the size of the seedset! ∎

This IP finds optimal connected activation sequences



$\theta = 2$
$\theta = 4$
$\theta = 8$
$\theta = 12$

Let $\mathbf{x_{it}}$ = **1** if node **i** activates at time **t**
0 otherwise

$$\min \sum_i \boxed{\sum_{t<\theta(i)} x_{it}}$$ (minimizes size of seedset)

= **1** if **i** is seed

Subject to:

$$\sum_t x_{it} = 1$$ (every node eventually activates)

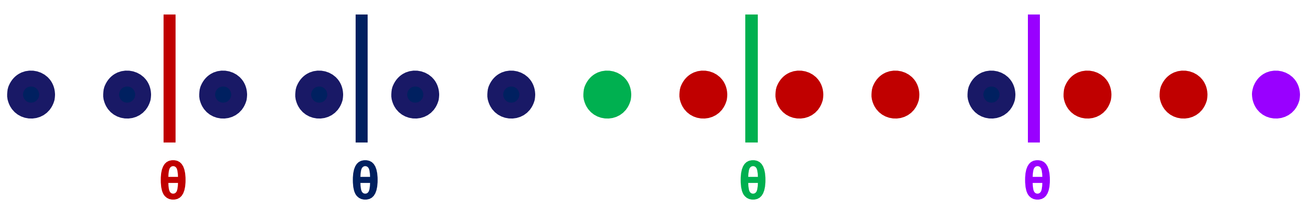
$$\sum_i x_{it} = 1$$ (one node activates per timestep)

$$\sum_{\text{edges }(i,j)} \boxed{\sum_{\tau<t} x_{j\tau}} \geq x_{it}$$ (connectivity)

= **1** if neighbor **j** is on by time **t**

Cor: IP returns seedset of size < **2opt.**

Activation sequence



Part II: How do we round this?

Iterative and adaptive rounding
with **both** the seedset and sequence.

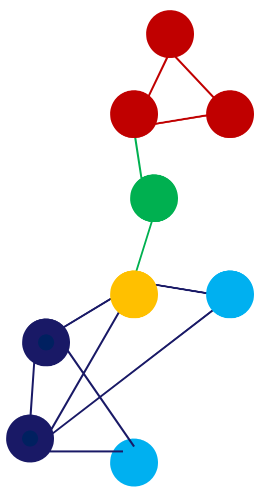
We return **connected seedsets**
instead of **connected activation sequences**.
(⇨ O(r)-approx instead of 2-approx)

Rounding the seedset or the sequence?

Because integer programs are not efficient, we relax the IP to a linear program (LP).

Now the $\mathbf{x}_{it}$ are fractional value on [0,1]. How can we round them to an integers?

$\theta = 1$
$\theta = 3$
$\theta = 4$
$\theta = 5$
$\theta = 7$



Optimal Seedset:

Threshold θ is ✅ if at least θ nodes are active by time θ

Approach 1: Sample the seedset.

$\mathbf{i}$ is a seed with probability $\propto \sum_{t<\theta(i)} \mathbf{x}_{it}$

Pro: Small seedset.

Con: No guarantee that every node activates.

Approach 2: Sample the activation sequence.

$\mathbf{i}$ activates by time $\mathbf{t}$ with probability $\propto \sum_{\tau<t} \mathbf{x}_{i\tau}$

Pro: Every node is activated.

Con: Corresponding seedset can be huge!

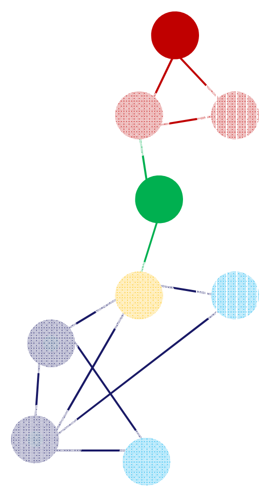
Solution?

Approach 3: Sample both together.

Then reconcile them adaptively & iteratively.

Approach 3: Sample seedset and sequence together!

$\theta = 1$
$\theta = 3$
$\theta = 4$
$\theta = 5$
$\theta = 7$

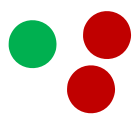


Sample seedset: (use Approach 1)

1. Let **i** be a seed with prob. $\mathbf{O(\log |V|)\, \Sigma_{t<\theta(i)}\, x_{it}}$
2. **Glue** seedset together so it's connected

This grows seedset by a factor of **O(r log |V|)**

Sampled seedset:



Construct an activation sequence deterministically:

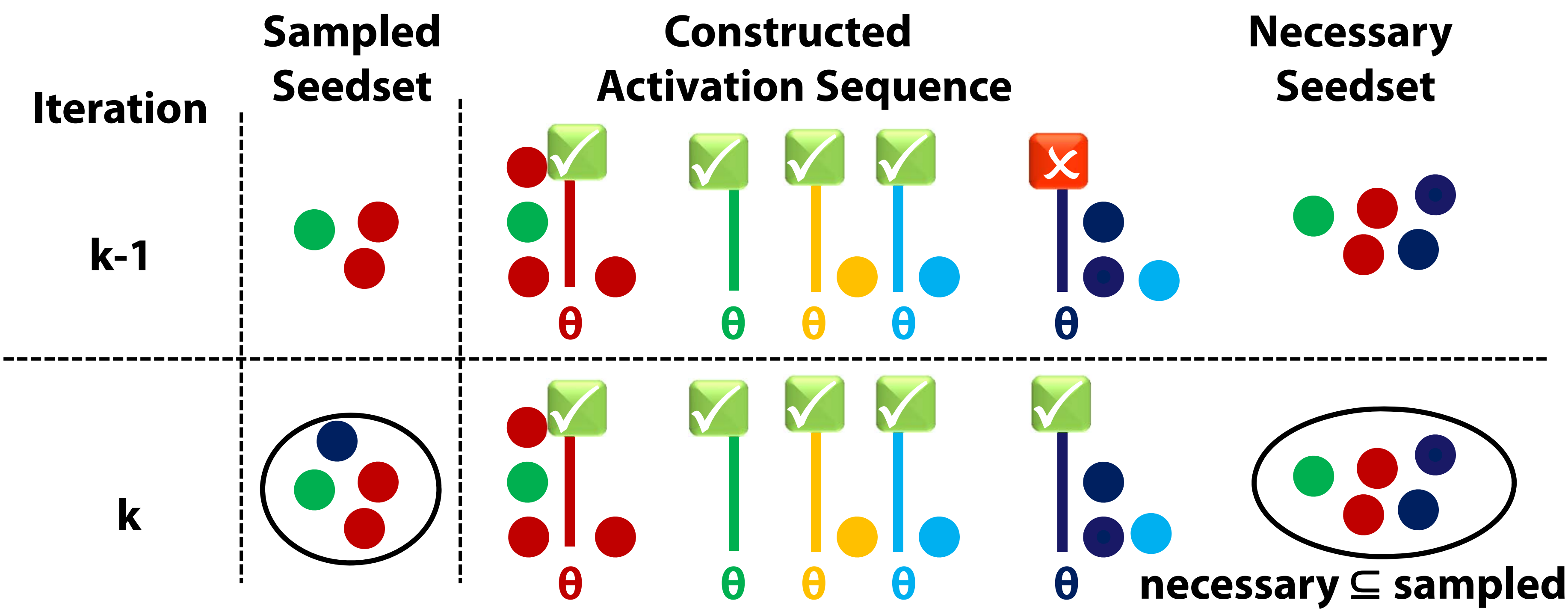
- Activate all the seeds at time **1**
- For each timestep **t**
  - For every inactive node connected to active node
  - … activate it if it has threshold $\boldsymbol{\theta > t}$

Constructed Activation Sequence:

Iteratively round both seedset and sequence!

At iteration j:

- Use rejection sampling to add **extra** nodes to sampled seedset
- … so that $\boldsymbol{\theta_j}$ is ✅ in constructed activation sequence.



When all **θ** are ✅, con

Threshold **θ** is ✅
if at least **θ** nodes
are active by time **θ**

By how much does this grow the seedset?
k thresholds, with **O(r log|V|)** increase per threshold.
Total **O(r k log|V|)** growth.

Why does this work?

How to show: For each iteration **j**, rejection sampling ensures

$\boldsymbol{\theta_j}$ is ✅ in constructed seedset?

Approach 3: Sample seedset.

- Let **i** be a seed with prob. $\propto \sum_{t<\theta(i)} x_{it}$

Deterministically construct sequence:

- Activate all the seeds at time **1**
- For each timestep **t**
  - Activate all nodes with $\boldsymbol{\theta > t}$
  - …that are connected to an active node

$\approx$

Approach 2: Sample the activation sequence.

- **i** activates by time **t** with probability $\propto \sum_{\tau<t} x_{i\tau}$

⇨ Enough nodes on by time $\mathbf{t = \theta_j}$, and $\boldsymbol{\theta_j}$ is ✅ !

With Approach 3 we gain:

1. Connectivity
2. Every node activates
3. Small seedset

This is the tricky part. Our proof uses two ideas:

Add **flow constraints** to LP

&

Activate seeds at **t=1** in constructed sequence.

(⇨ connected seedset)

Wrapping up

Minimization formulation: Given the graph and thresholds **θ**, find the smallest seedset that activates every node in the graph.

Main result: An $O(r \cdot k \cdot \log |V|)$-approx algorithm based on LPs

r is graph diameter, **k** is number of possible thresholds

Algorithm finds **connected seedsets**.

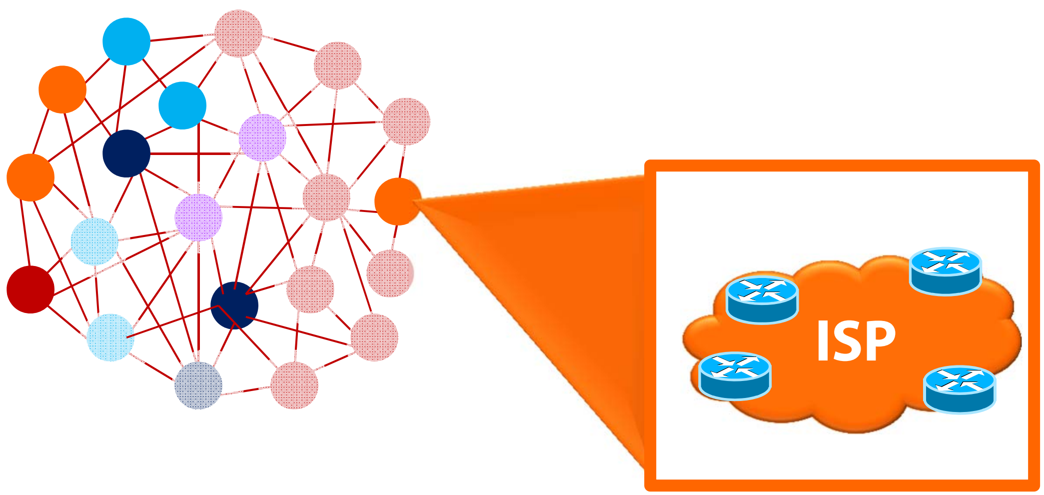
Lower Bound: Can't do better than an $\Omega(\log |V|)$ approx. (Even for constant **r, k**)

Lower Bound: Can't do better that an $\Omega(r)$ approx if seedset is connected.

Open problems:

- Can we solve without LPs?
- Can we gain something with random thresholds?
- Apply techniques in less stylized models? (e.g. models of Internet routing.)
- …

Thanks!



http://arxiv.org/abs/1202.2928