Network Cascades for Spatial Conservation Planning

Presenter: Bistra Dilkina

CS 6702

Feb 15, 2011

(most slides courtesy of Dan Sheldon)

Overview

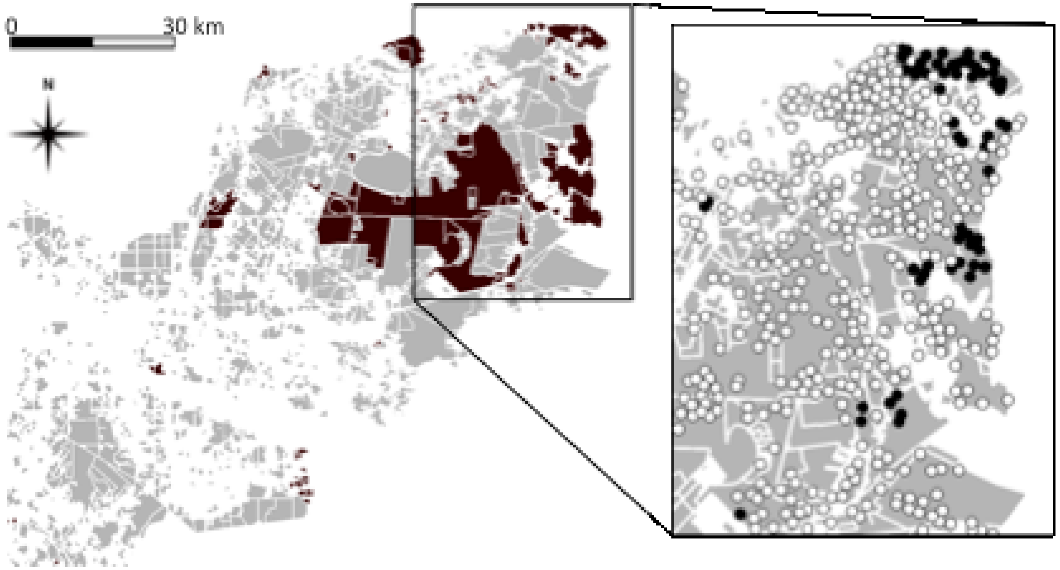
- Collaborative project to develop optimal conservation strategies for Red-Cockaded Woodpecker (RCW)
  - Institute for Computational Sustainability (Cornell and OSU): Daniel Sheldon, Bistra Dilkina, Adam Elmachtoub, Ryan Finseth, Ashish Sabharwal, Jon Conrad, Carla P. Gomes, David Shmoys
  - The Conservation Fund: Will Allen, Ole Amundsen, Buck Vaughan
- Recent paper: Maximizing the Spread of Cascades Using Network Design, UAI 2010
- Key Idea: *RCW population dynamics as a network cascade*

Part I: Problem Setup

RCW population dynamics as a network cascade

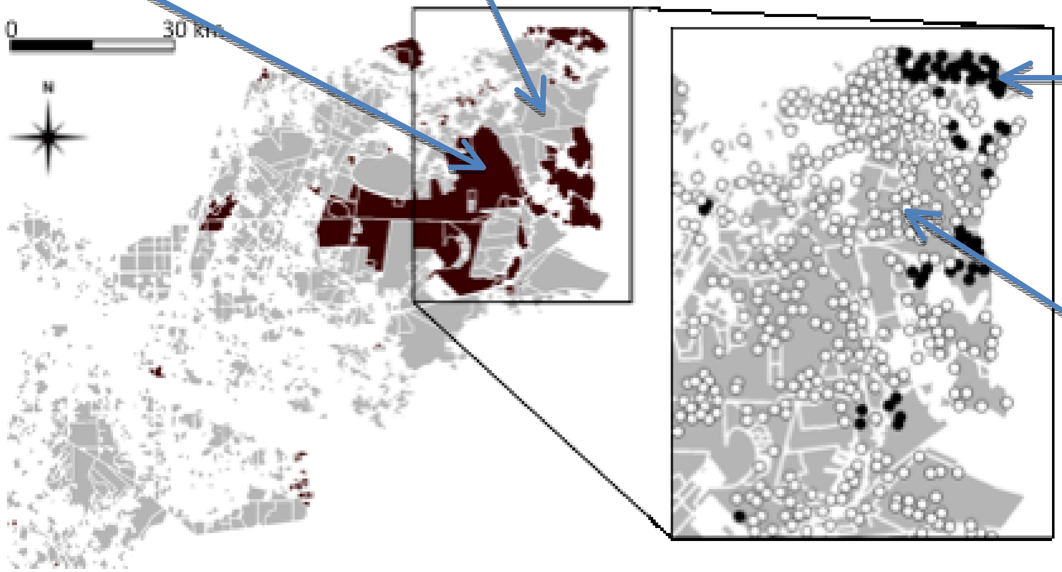
Application:
Spatial Conservation Planning

- What is the best land acquisition and management strategy to support the recovery of the Red-Cockaded Woodpecker (RCW)?



Federally listed rare and endangered species

Problem Setup



Conserved parcels

Available parcels

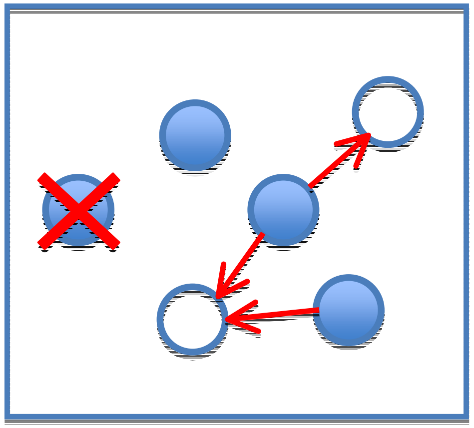
Current territories

Potential territories

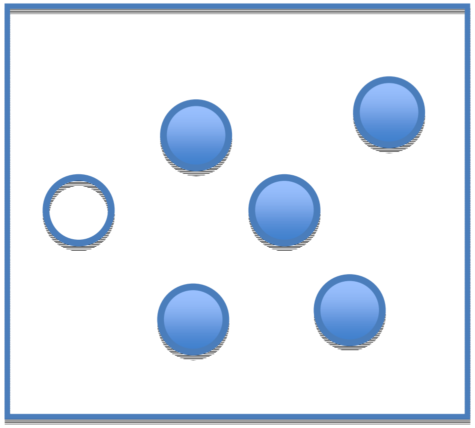
Given limited budget, what parcels should I conserve to maximize the expected number of occupied territories in 50 years?

Metapopulation Model

- Model for population dynamics in fragmented landscape
  - Territories are occupied or unoccupied in each time step
  - Two types of stochastic events:
    - *Local extinction:* occupied -> unoccupied
    - *Colonization:* unoccupied -> occupied (from neighbor)



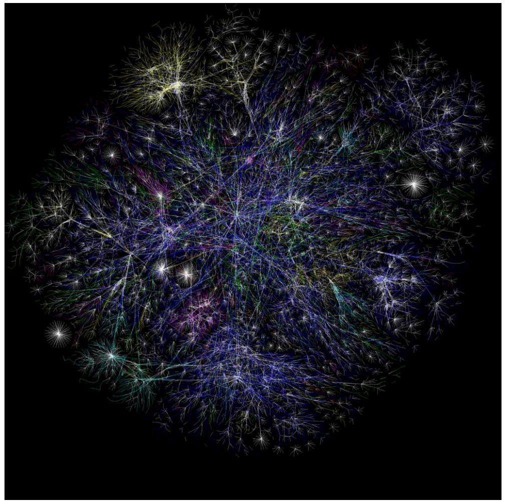
Time 1

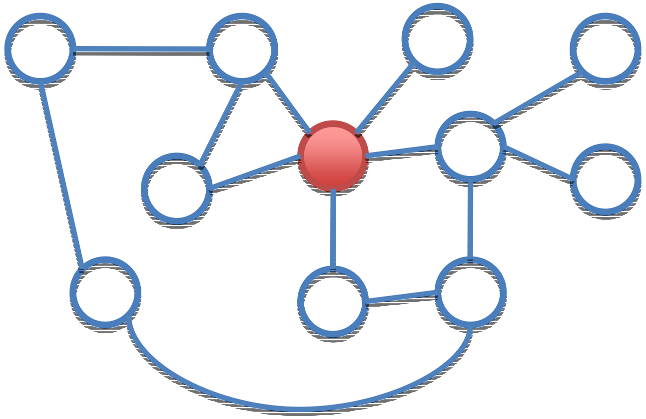


Time 2

Network Cascades

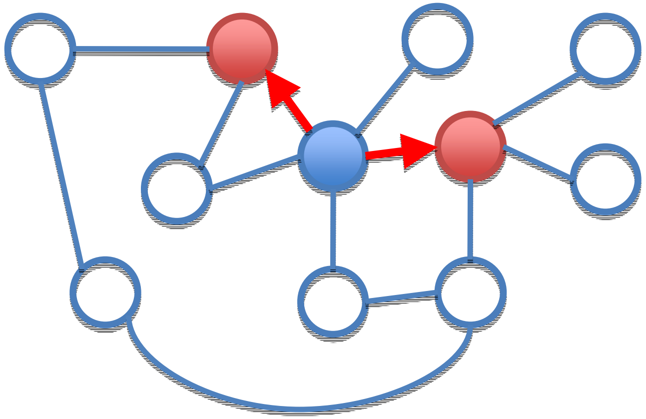
- Models for diffusion in (social) networks
  - Spread of information, behavior, disease, etc.
  - E.g.: suppose each individual passes rumor to friends independently with probability ½





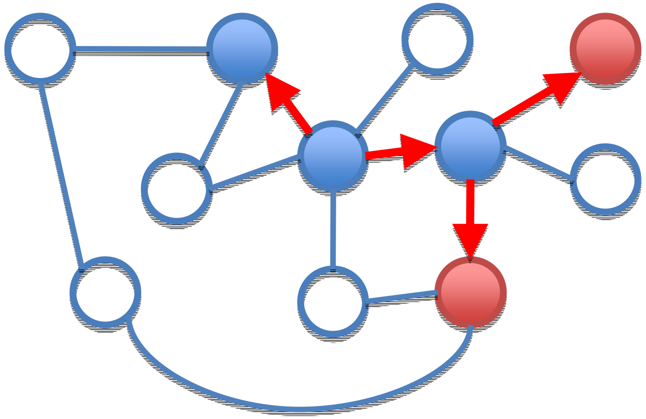
Network Cascades

- Models for diffusion in (social) networks
  - Spread of information, behavior, disease, etc.
  - E.g.: suppose each individual passes rumor to friends independently with probability ½



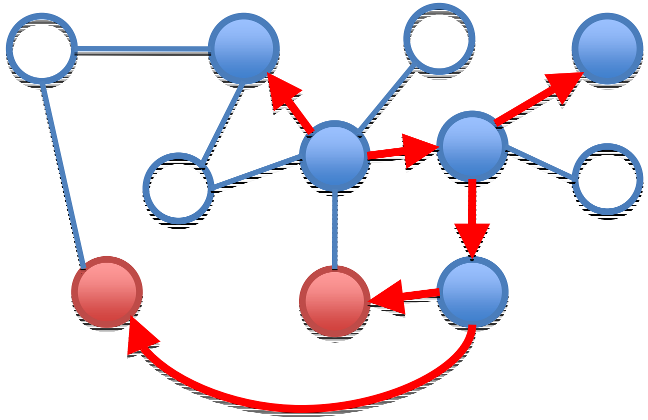
Network Cascades

- Models for diffusion in (social) networks
  - Spread of information, behavior, disease, etc.
  - E.g.: suppose each individual passes rumor to friends independently with probability ½



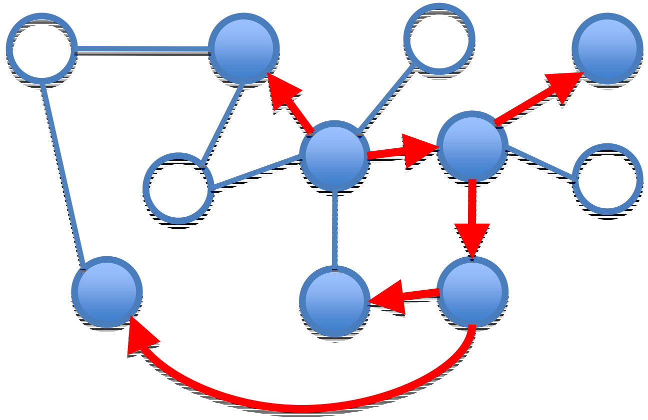
Network Cascades

- Models for diffusion in (social) networks
  - Spread of information, behavior, disease, etc.
  - E.g.: suppose each individual passes rumor to friends independently with probability ½



Network Cascades

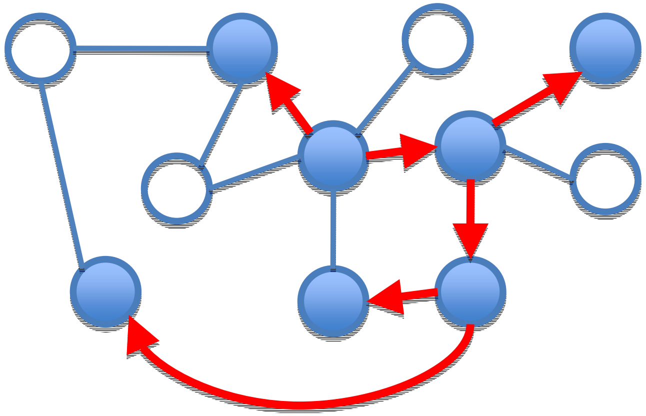
- Models for diffusion in (social) networks
  - Spread of information, behavior, disease, etc.
  - E.g.: suppose each individual passes rumor to friends independently with probability ½



Note: “activated” nodes are those reachable by red edges

Optimization

- We often want to *intervene* to achieve some goal: optimization!

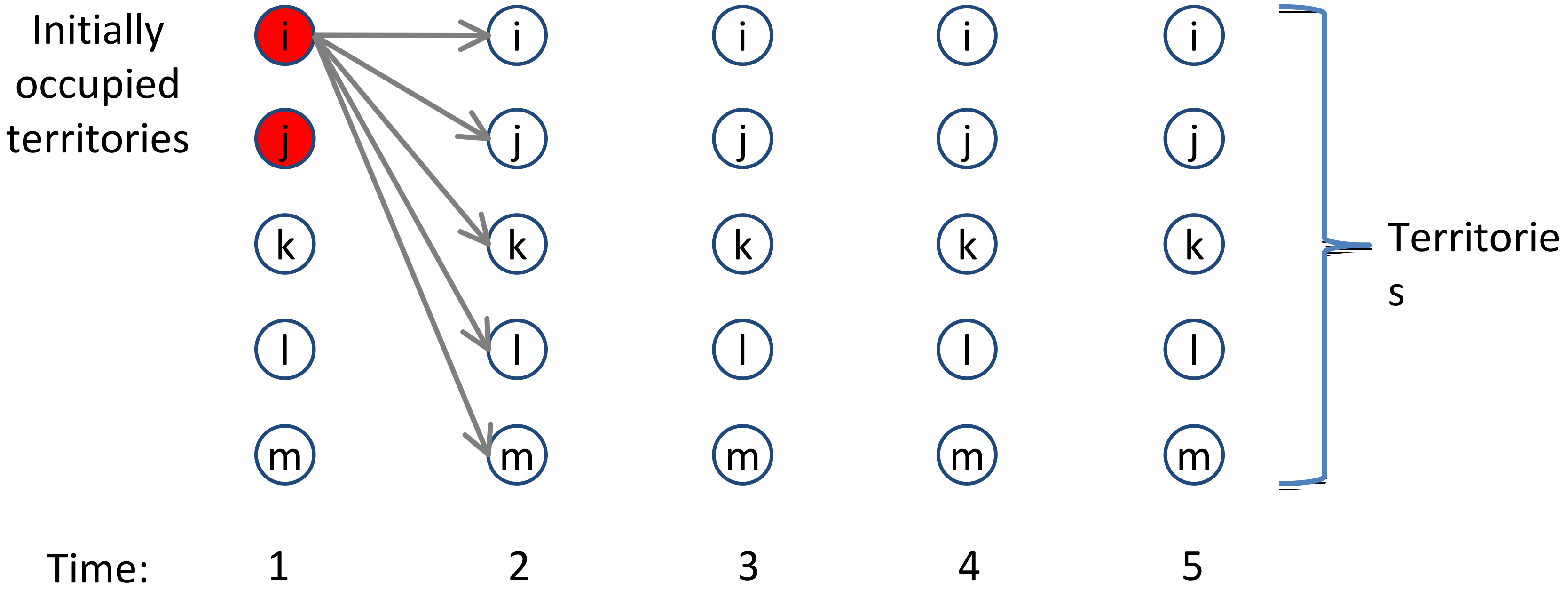


- Viral marketing: which customers to target to maximize the effectiveness of word-of-mouth marketing?

[Domingos and Richardson, KDD, 2001], [Kempe, Kleinberg, Tardos, KDD, 2003]

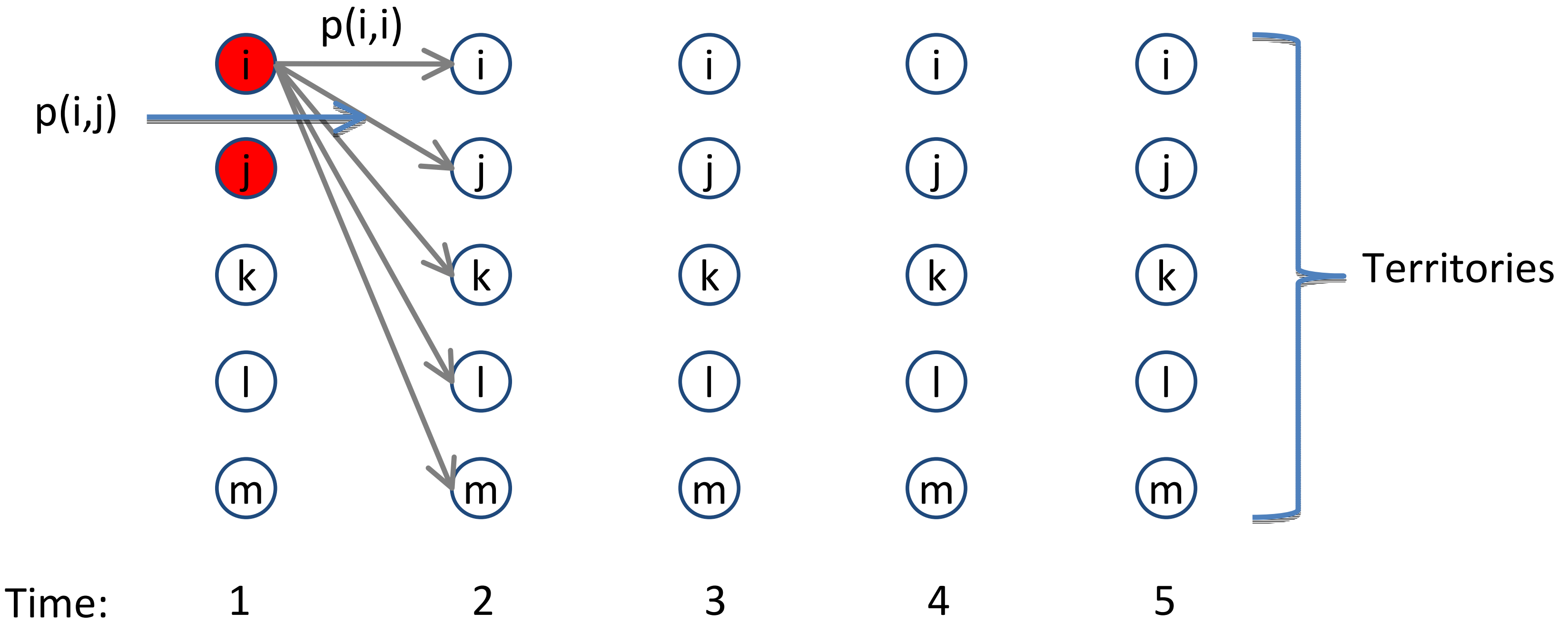
Metapopulation = Cascade

- Metapopulation model can be viewed as a cascade in the *layered* graph representing territories over time



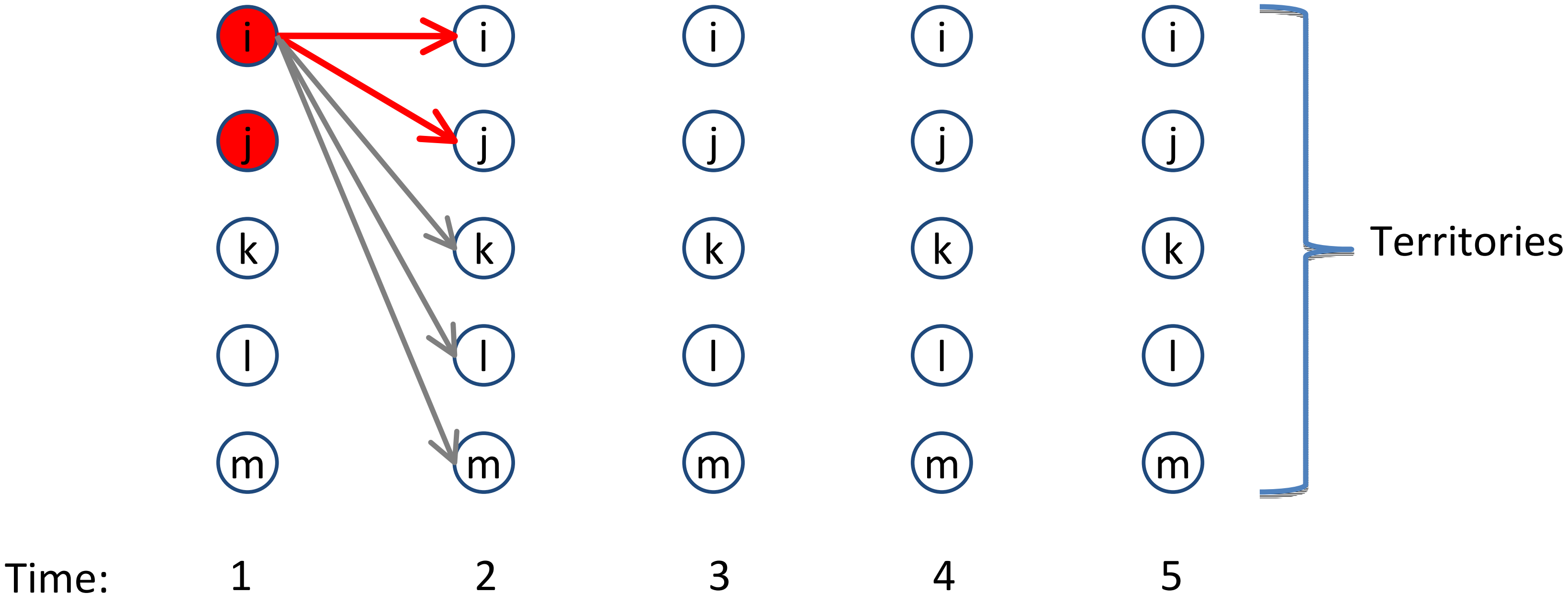
Metapopulation = Cascade

- Metapopulation model can be viewed as a cascade in the *layered* graph representing territories over time



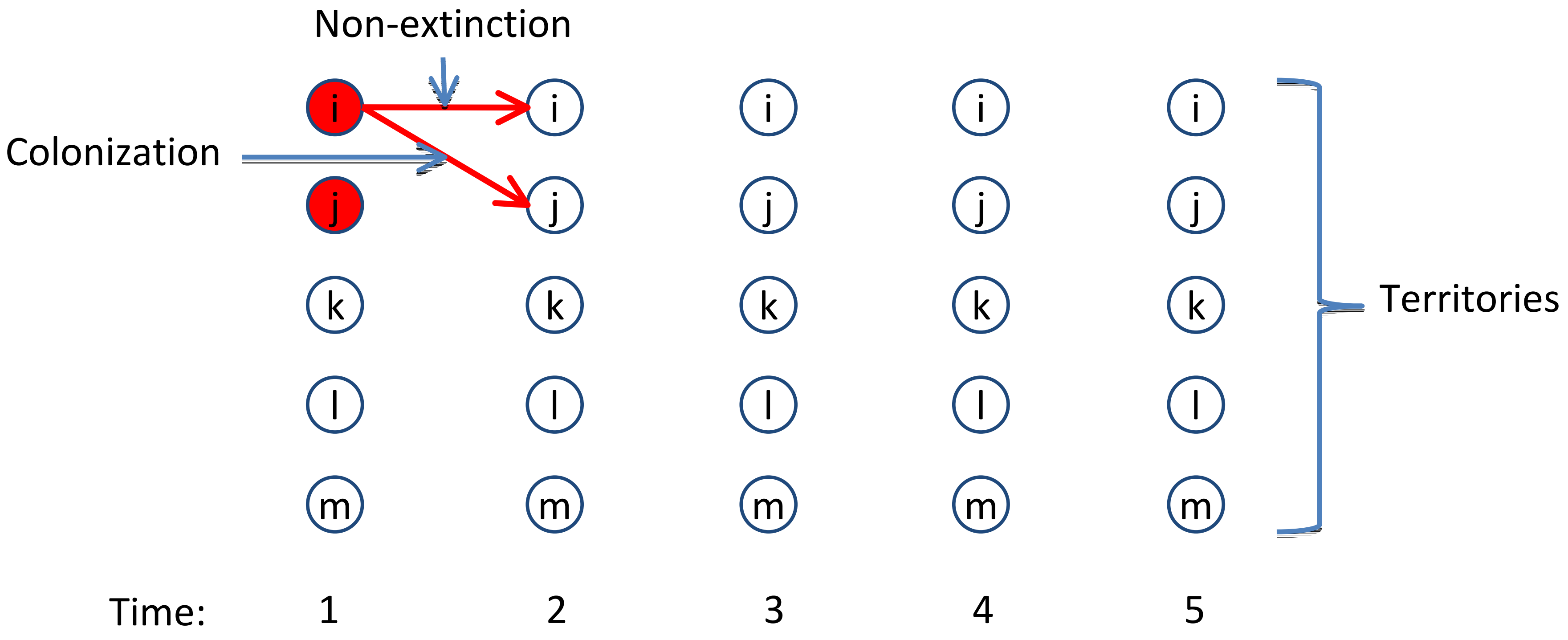
Metapopulation = Cascade

- Metapopulation model can be viewed as a cascade in the *layered* graph representing territories over time



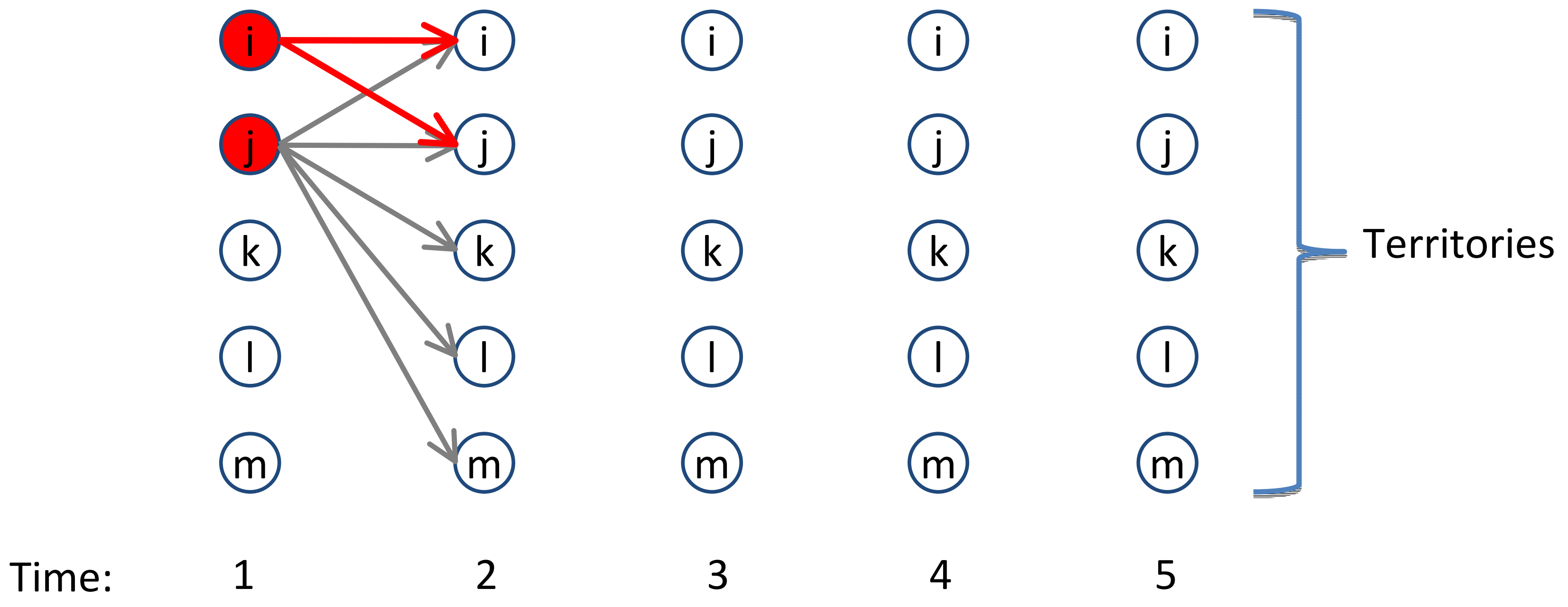
Metapopulation = Cascade

- Metapopulation model can be viewed as a cascade in the *layered* graph representing territories over time



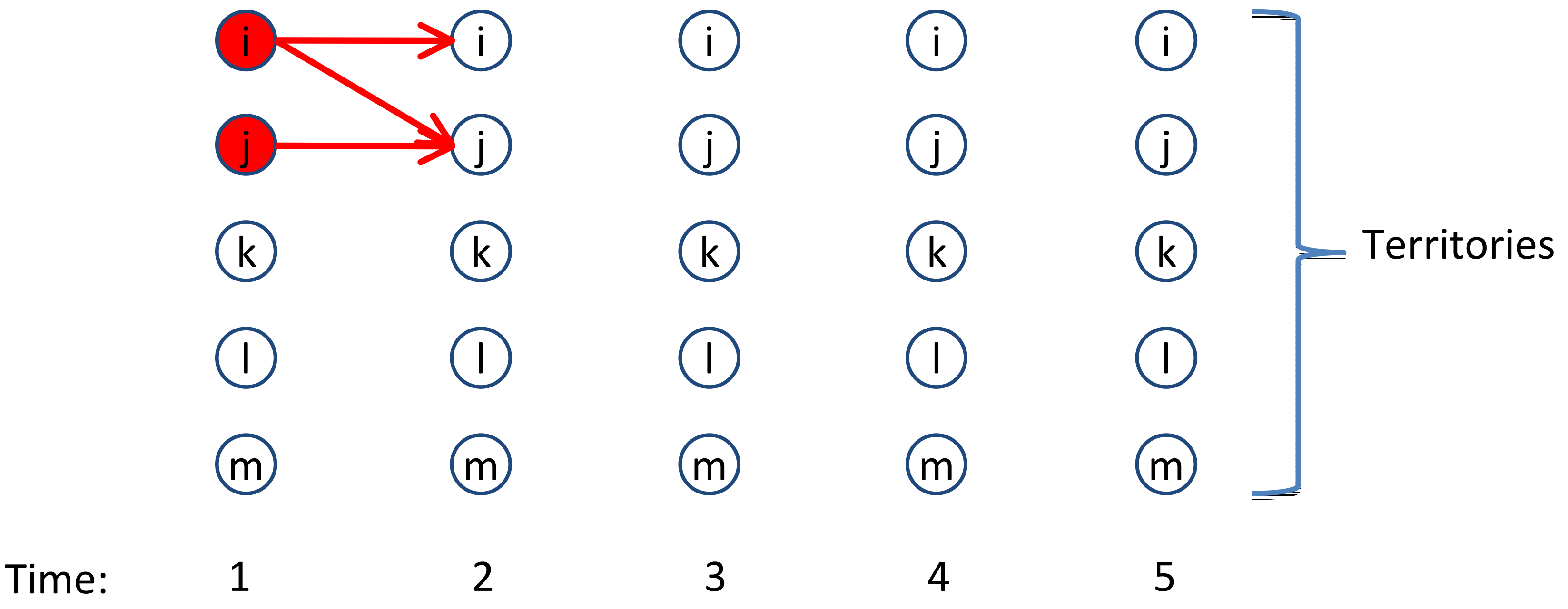
Metapopulation = Cascade

- Metapopulation model can be viewed as a cascade in the *layered* graph representing territories over time



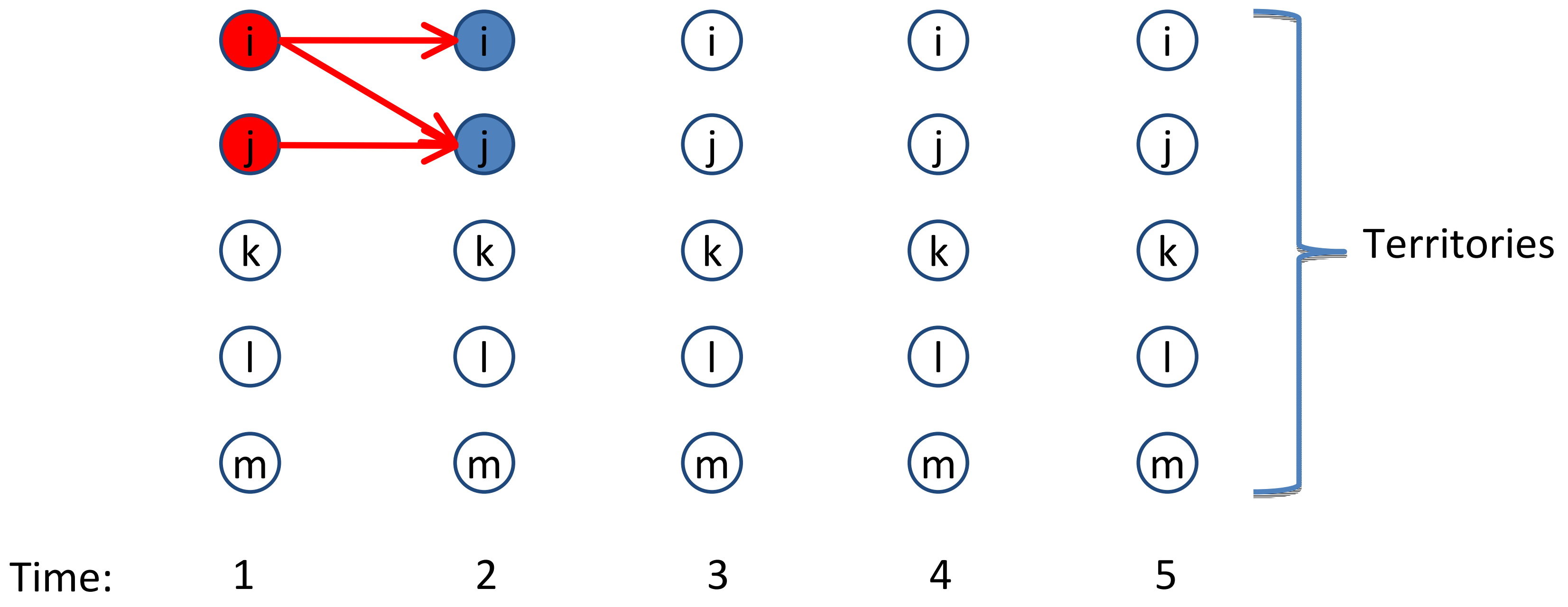
Metapopulation = Cascade

- Metapopulation model can be viewed as a cascade in the *layered* graph representing territories over time



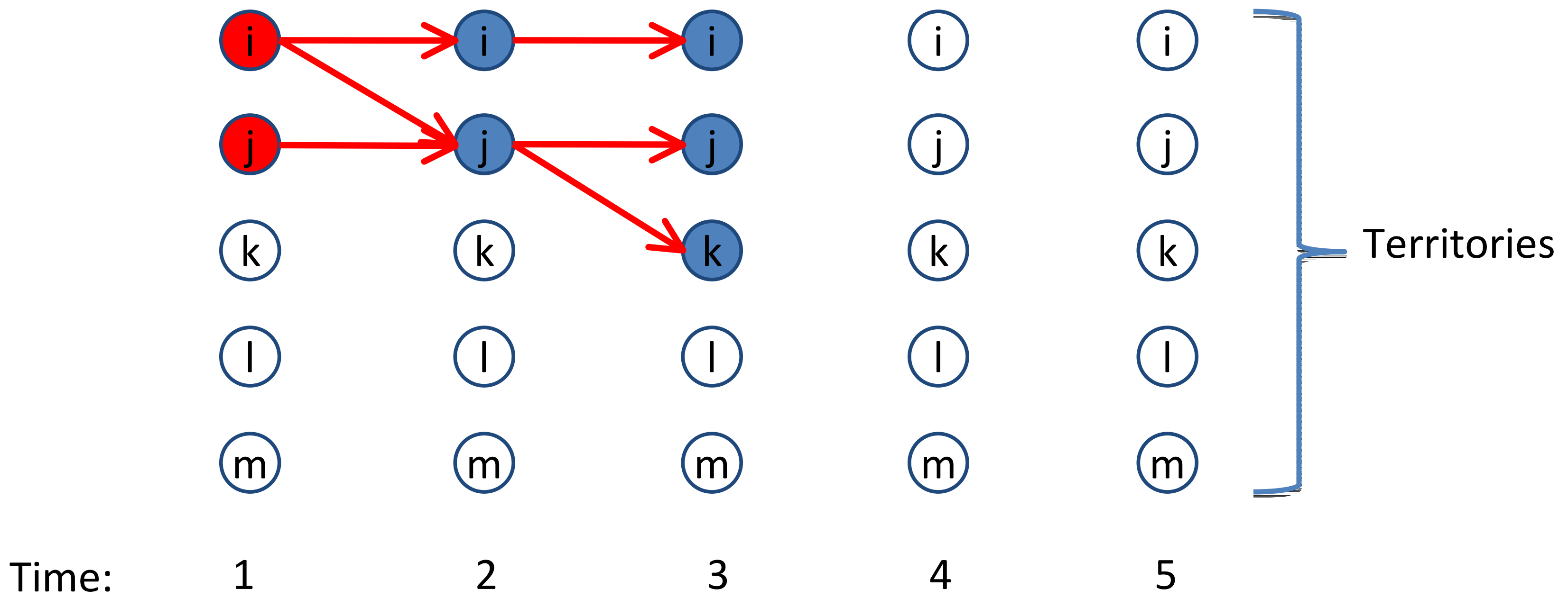
Metapopulation = Cascade

- Metapopulation model can be viewed as a cascade in the *layered* graph representing territories over time



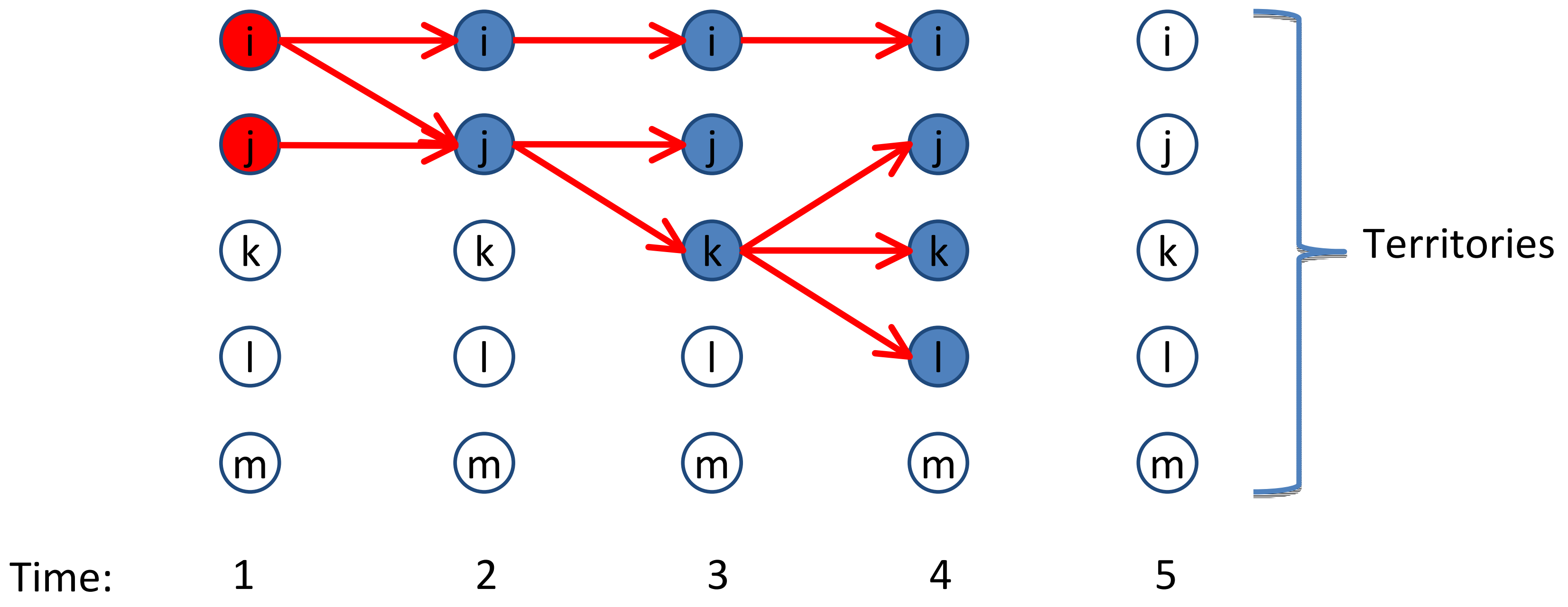
Metapopulation = Cascade

- Metapopulation model can be viewed as a cascade in the *layered* graph representing territories over time



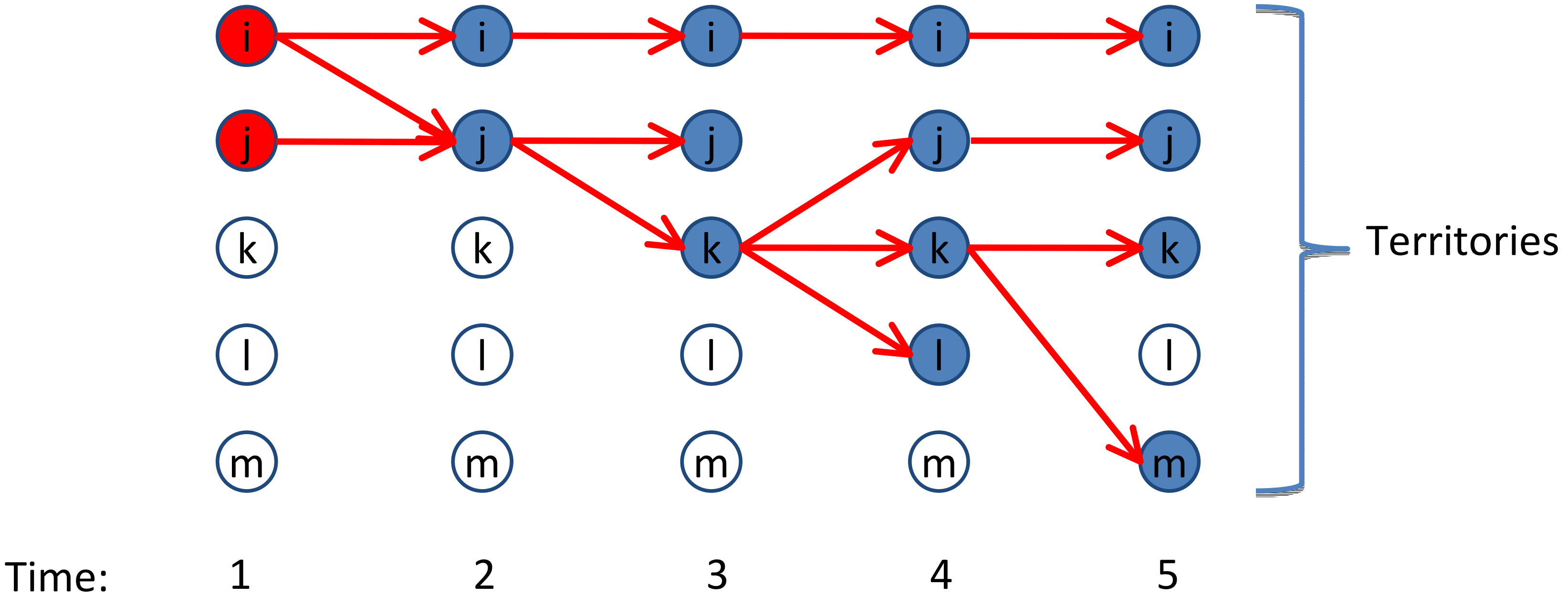
Metapopulation = Cascade

- Metapopulation model can be viewed as a cascade in the *layered* graph representing territories over time



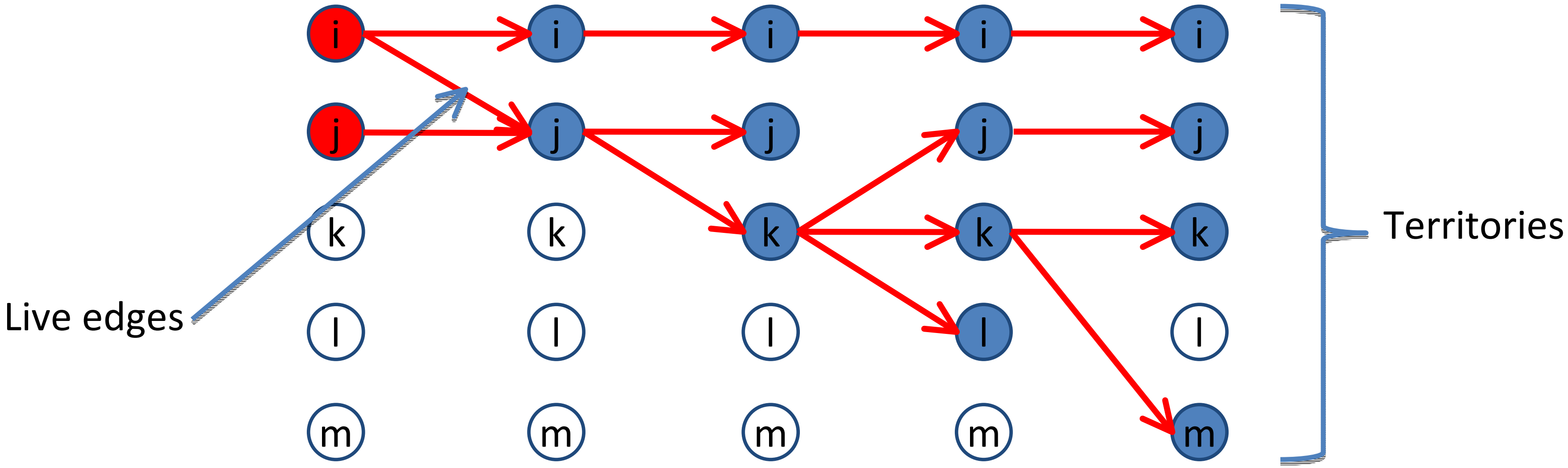
Metapopulation = Cascade

- Metapopulation model can be viewed as a cascade in the *layered* graph representing territories over time



Metapopulation = Cascade

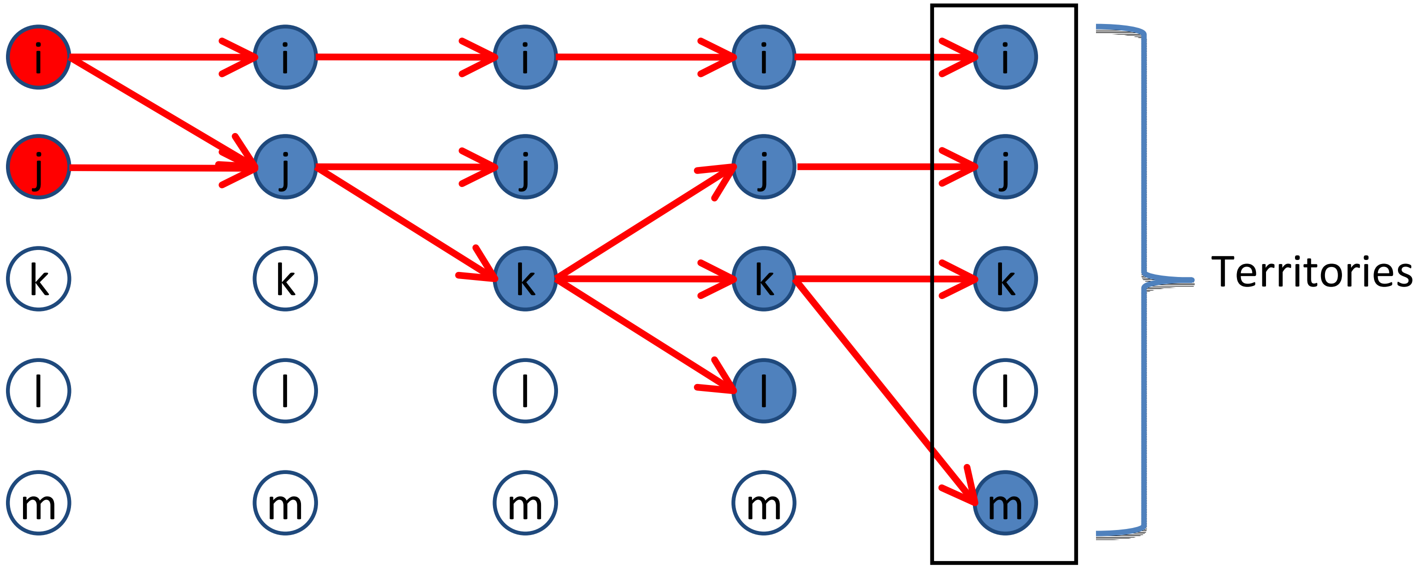
- Metapopulation model can be viewed as a cascade in the *layered* graph representing territories over time



Key point: after simulation, occupied territories given by nodes that are *reachable in the network* by live edges

Metapopulation = Cascade

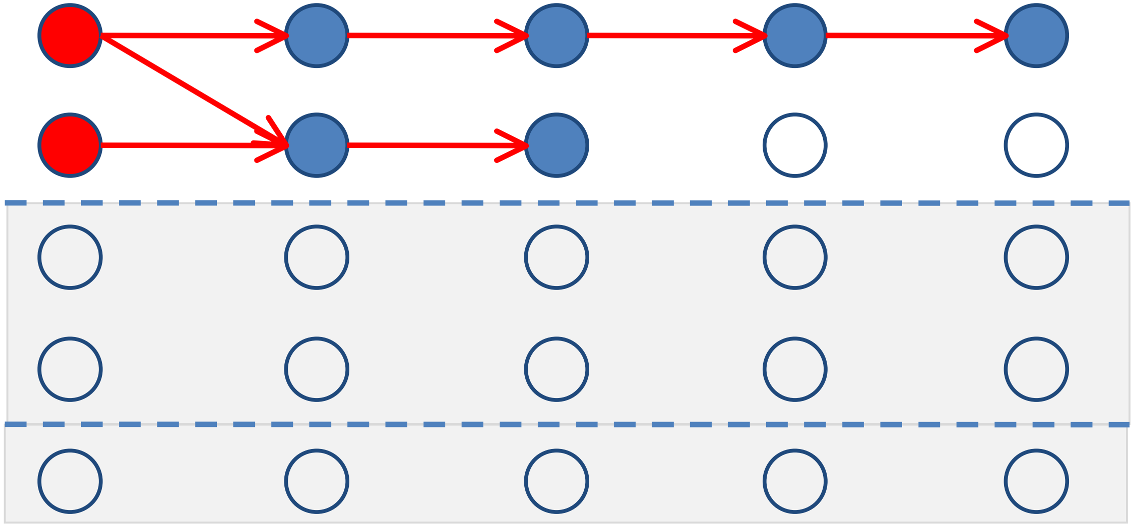
- Metapopulation model can be viewed as a cascade in the *layered* graph representing territories over time



Target nodes: territories at final time step

Management Actions

- Conserving parcels adds nodes to the network to create new pathways for the cascade



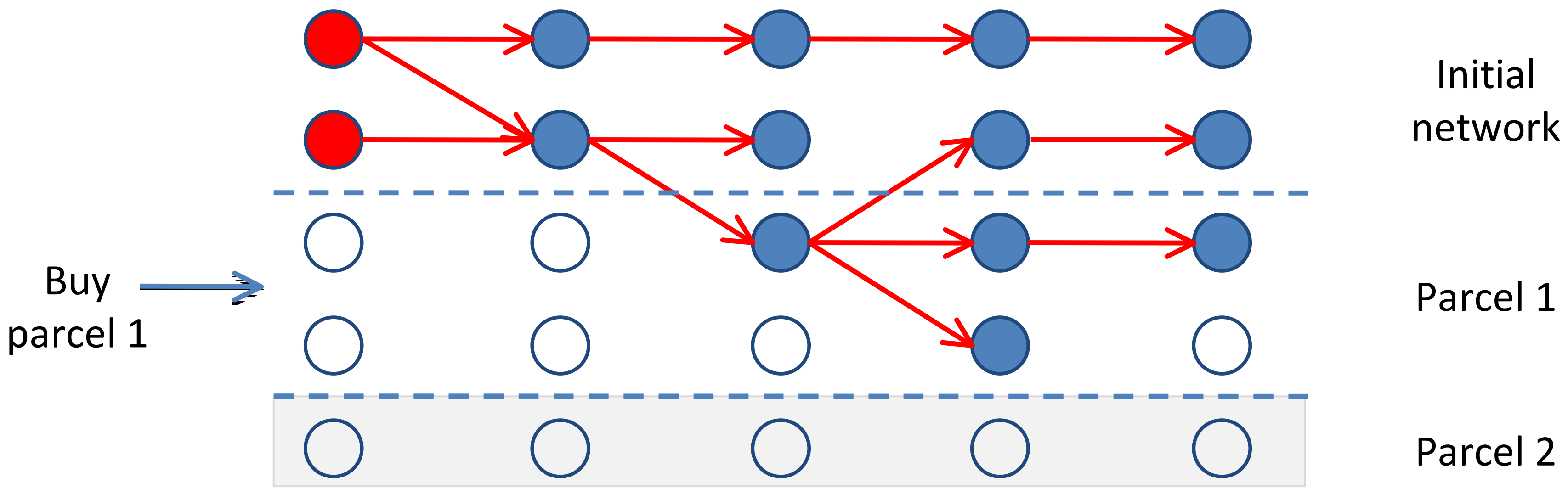
Initial network

Parcel 1

Parcel 2

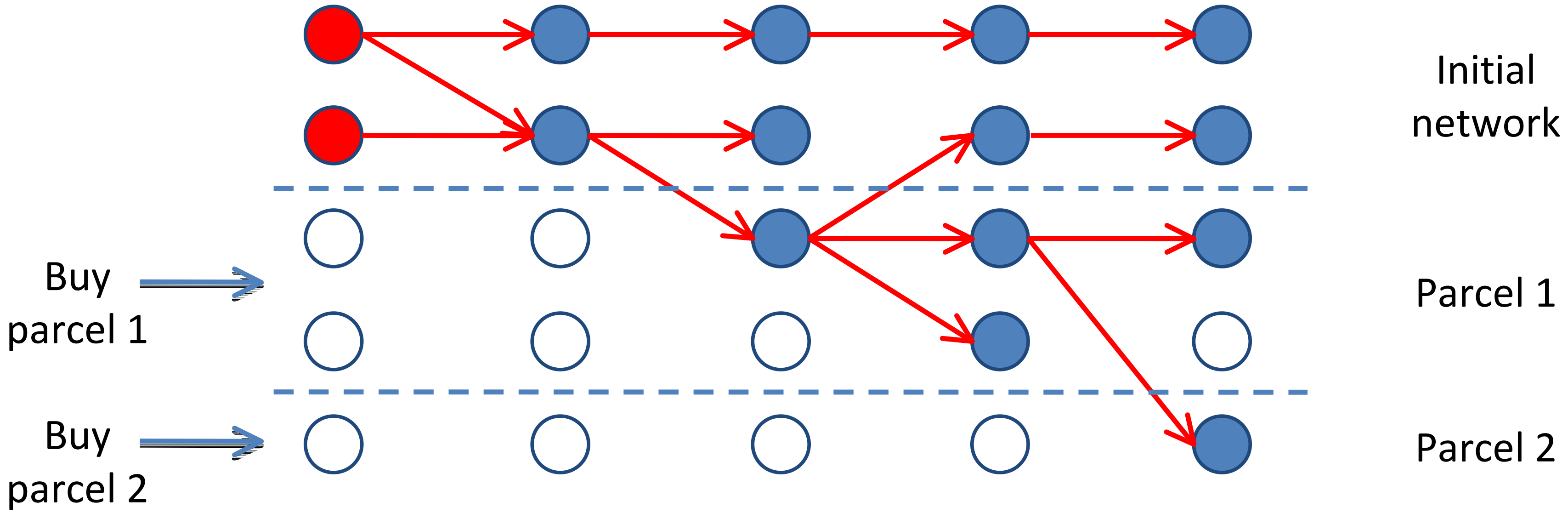
Management Actions

- Conserving parcels adds nodes to the network to create new pathways for the cascade



Management Actions

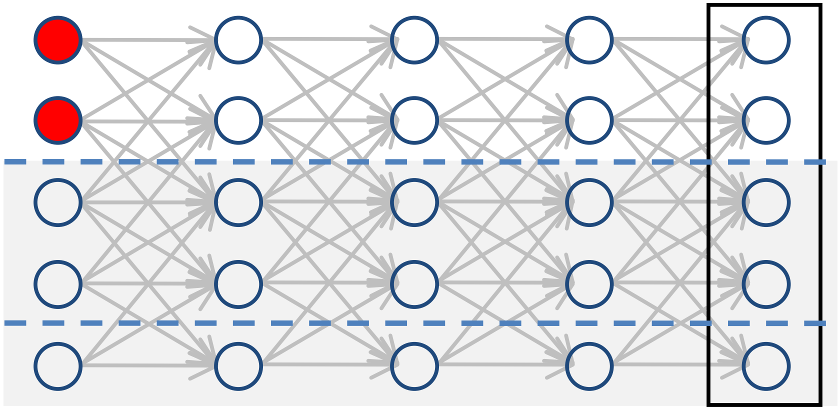
- Conserving parcels adds nodes to the network to create new pathways for the cascade



Cascade Optimization Problem

Given:

- Territory network with colonization and extinction probabilities
- Initially occupied territories
- Initial network
  - parcels that are already conserved
- Management actions
  - list of available parcels and their costs
- Time horizon *T*
- Budget *B*



Find set of parcels with total cost at most B that maximizes the ***expected*** *number of occupied territories at time T.*

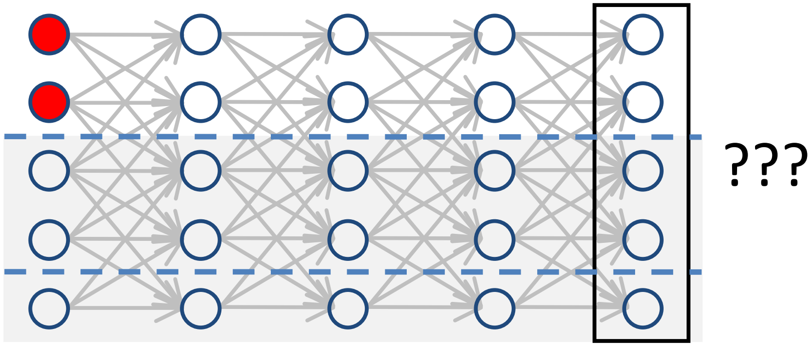
Notes

- A *stochastic* optimization problem
  - We don’t know exactly what the result of our actions will be...
  - Have a probability model (metapopulation)
  - Goal: optimize *expected value* of a random variable
- Questions on problem formulation?

Part II: Optimization

Stochastic Optimization

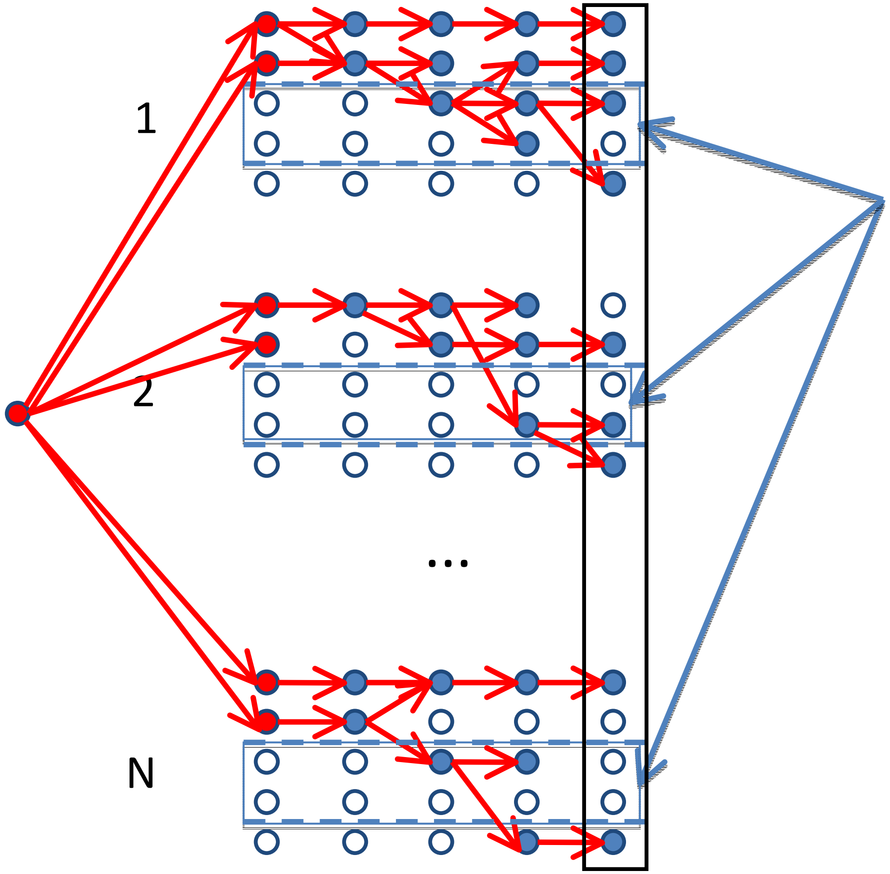
- Stochastic problem is unwieldy: we cannot even calculate the probabilities required to compute objective (#P-hard)



- Sample average approximation (SAA):
  - Replace stochastic problem by deterministic analog
  - Draw *N* outcomes from underlying probability space and optimize empirical average instead of expectation

Sample Average Approximation

- Sample *N* *training cascades* by flipping coins for all edges.



- Select single set of management actions that works well “in hindsight” for training cascades.
- Goal: maximize the number of reachable target nodes
- A *deterministic network design problem*. Can leverage existing techniques to formulate and solve as mixed integer program

Sample Average Approximation

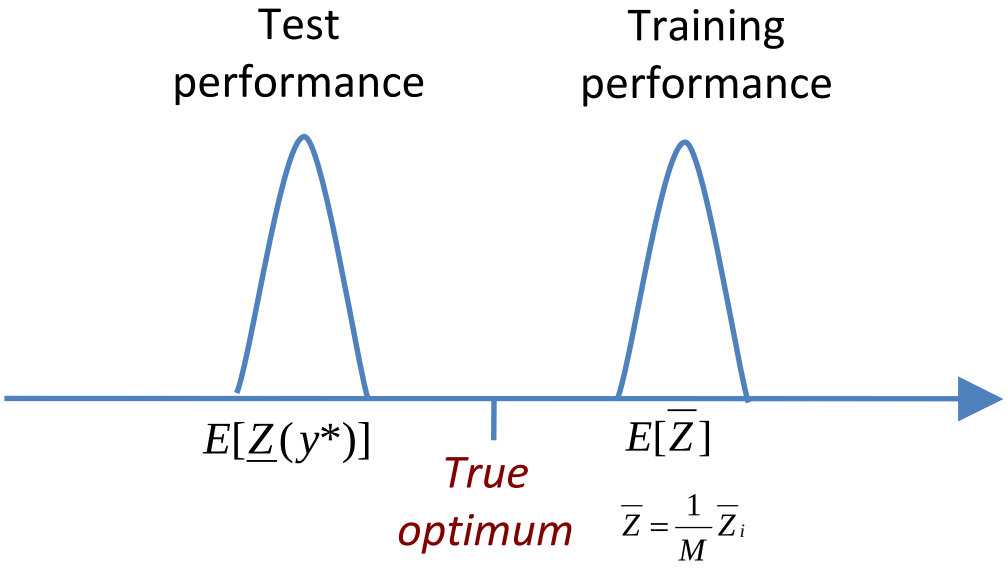
- Repeat *M* times for i=1..M
  - Sample *N* *training cascades* by flipping coins for all edges.
  - Solve deterministic optimization problem to obtain buying strategy y_i with optimum *training objective* $\overline{Z}_i$ (empirical average over the *N* cascades)
  - Evaluate buying strategy an a large sample of *Nvalid* *validations cascades* and record *validation objective* (empirical average over *Nvalid* cascades)
- Choose the best buying strategy y^* among the *M* proposed strategy according to validation objective
- Evaluate best buying strategy an a large sample of *Ntest* *test cascades* and record *test objective* $\underline{Z}(y^*)$ (empirical average over *Ntest* cascades)

Wait a minute...

- The SAA is an *approximation* of the stochastic problem we want to solve
- How good is it?

SAA Bounds

- Can derive statistical estimates of upper and lower bound of true optimum
- Can reduce variance to get confident bounds by increasing *M* and *Ntest*



Test performance

Training performance

$E[\underline{Z}(y^*)]$

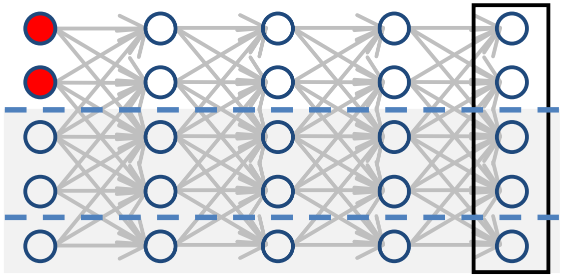
True optimum

$E[\overline{Z}]$

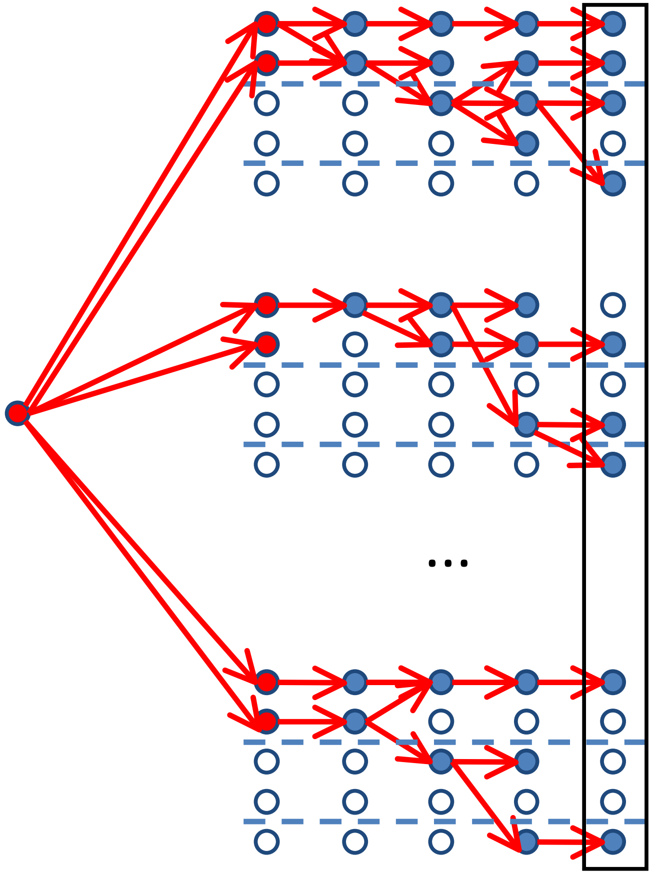
$\overline{Z} = \frac{1}{M}\overline{Z}_i$

[Norkin et al 1998, Mak et al 1999, Kleywegt et al 2001]

Network Design



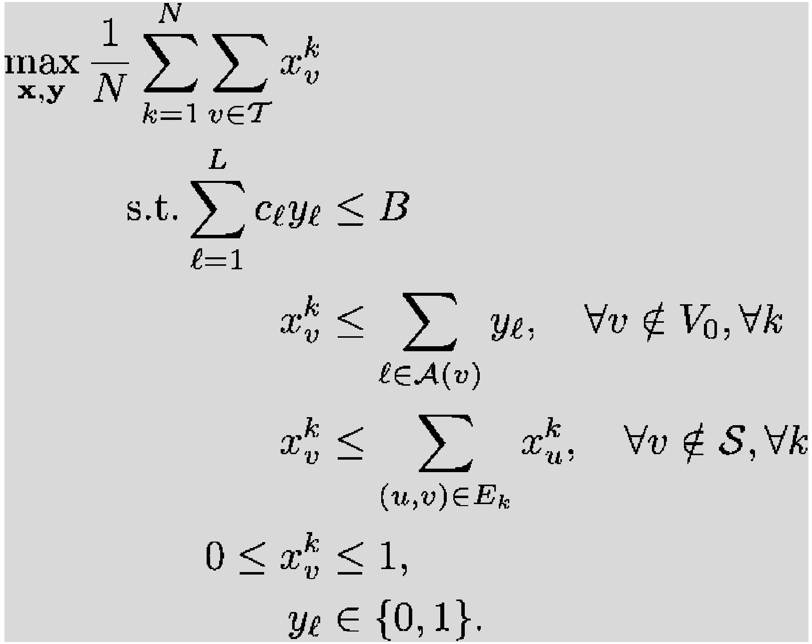
Stochastic



Deterministic over
N cascades

Network Design Mixed Integer Program

- Binary variables ***y*** to decide which actions to take
  - y_l = 1 if take action *l*, else 0
- Auxiliary variables ***x*** to encode reachability
- Set of constraints to enforce consistency between ***x*** and ***y***

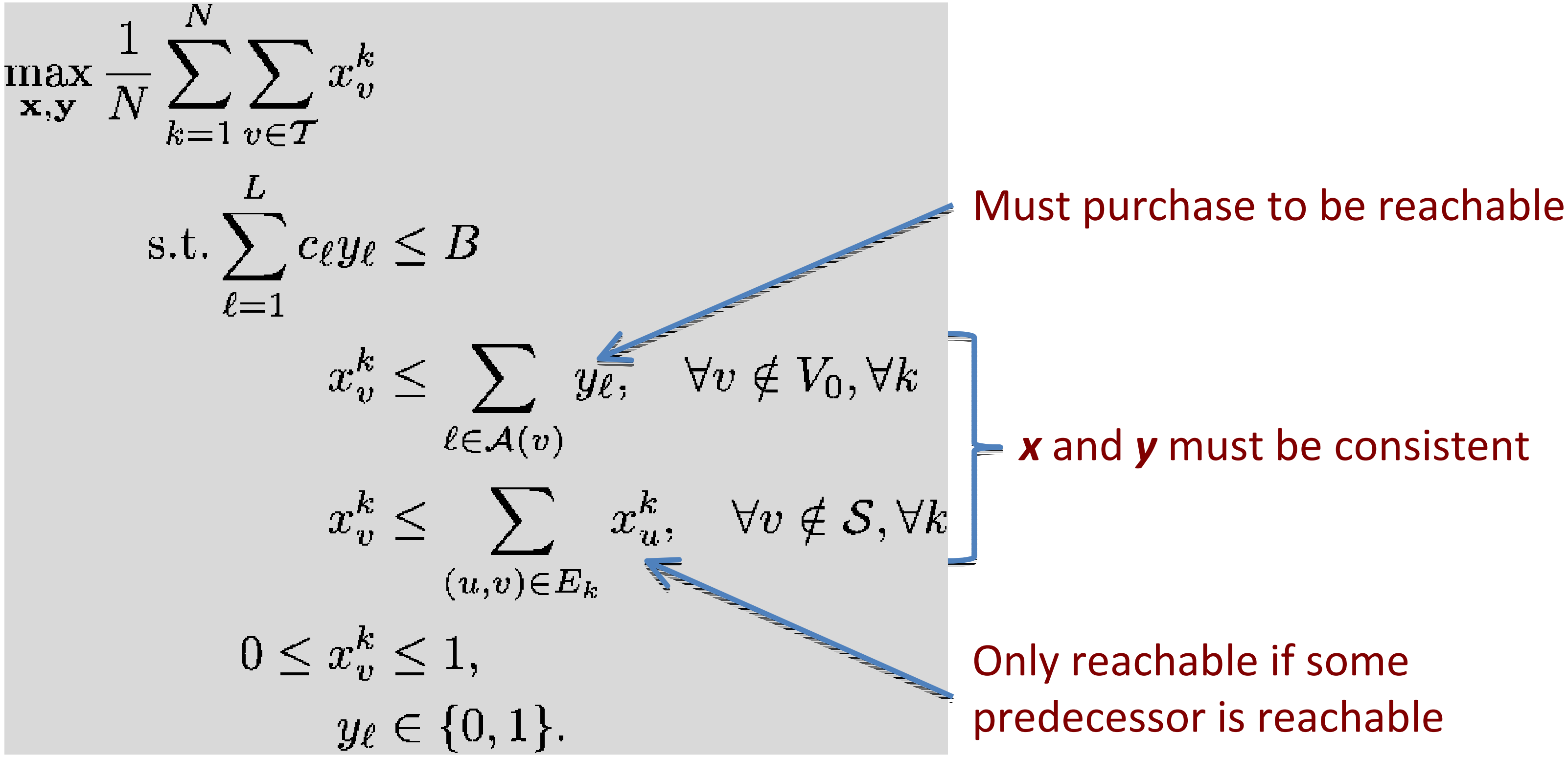


$$
\begin{aligned}
\max_{\mathbf{x},\mathbf{y}} \ & \frac{1}{N}\sum_{k=1}^{N}\sum_{v\in\mathcal{T}} x_v^k \\
\text{s.t.} \ & \sum_{\ell=1}^{L} c_\ell y_\ell \le B \\
& x_v^k \le \sum_{\ell\in\mathcal{A}(v)} y_\ell, \quad \forall v \notin V_0, \forall k \\
& x_v^k \le \sum_{(u,v)\in E_k} x_u^k, \quad \forall v \notin \mathcal{S}, \forall k \\
& 0 \le x_v^k \le 1, \\
& y_\ell \in \{0,1\}.
\end{aligned}
$$

NP hard: solve by branch and bound (CPLEX)

Network Design Mixed Integer Program

- Integer variables: $y_l = 1$ if take action l, else 0
- Introduce ***x*** variables to encode reachability, and add constraints to enforce consistency among ***x*** and ***y***



$$
\begin{aligned}
\max_{\mathbf{x},\mathbf{y}} \frac{1}{N} \sum_{k=1}^{N} \sum_{v \in \mathcal{T}} x_v^k \\
\text{s.t.} \sum_{\ell=1}^{L} c_\ell y_\ell \leq B \\
x_v^k \leq \sum_{\ell \in \mathcal{A}(v)} y_\ell, \quad \forall v \notin V_0, \forall k \\
x_v^k \leq \sum_{(u,v) \in E_k} x_u^k, \quad \forall v \notin \mathcal{S}, \forall k \\
0 \leq x_v^k \leq 1, \\
y_\ell \in \{0,1\}.
\end{aligned}
$$

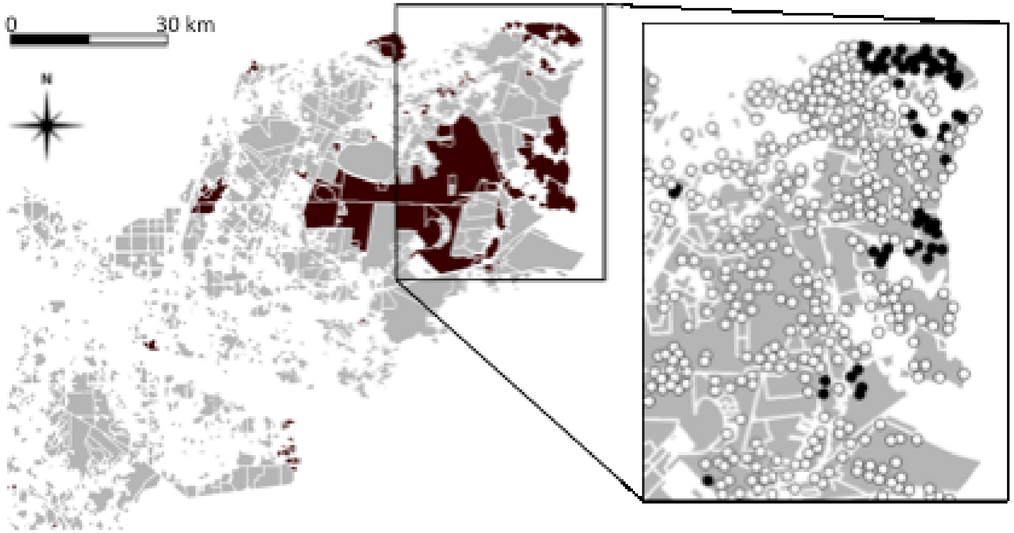
Must purchase to be reachable

x and ***y*** must be consistent

Only reachable if some predecessor is reachable

Experiments

- 443 available parcels
- 2500 territories
- 63 initially occupied
- 100 years



- Population model is parameterized based (loosely) on RCW ecology
  - Short-range colonizations (<3km) within the foraging radius of the RCW are much more likely than long-range colonizations

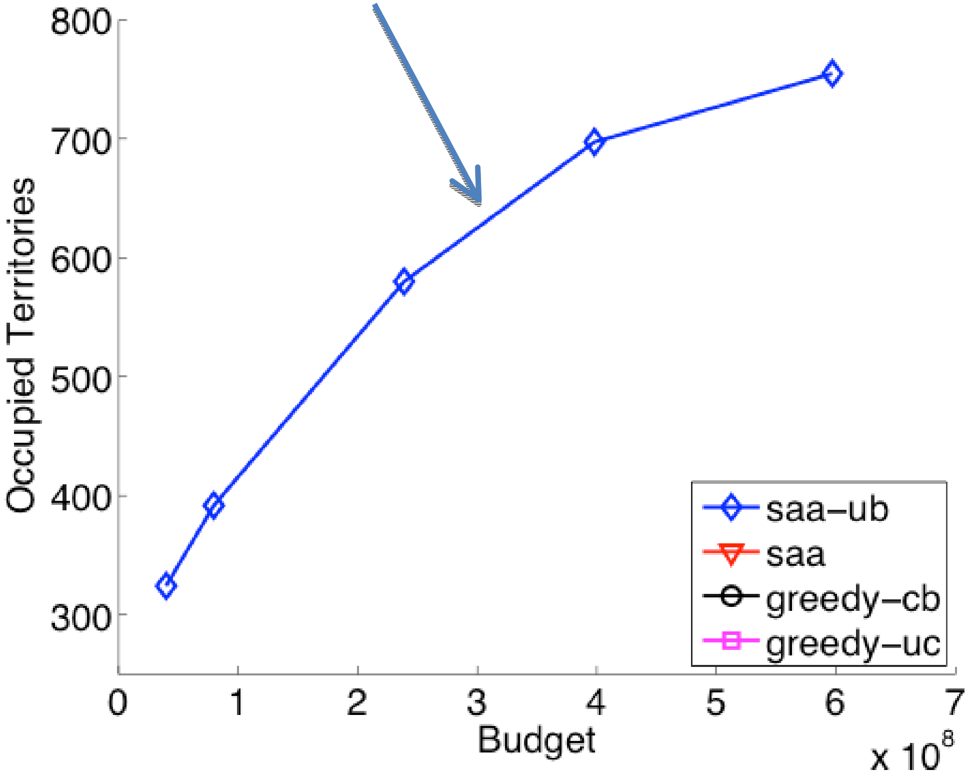
Greedy Baselines

- Adapted from previous work on influence maximization

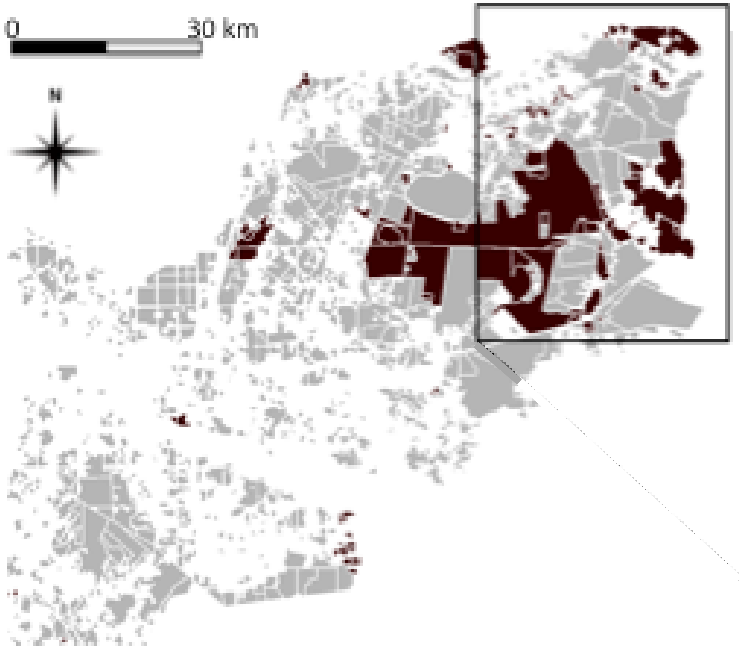
- Start with empty set, add actions until exhaust budget
  - **Greedy-uc** – choose action that results in biggest immediate increase in objective [Kempe et al. 2003]
  - **Greedy-cb** – use ratio of benefit to cost [Leskovec et al. 2007]

- No performance guarantees!

Results

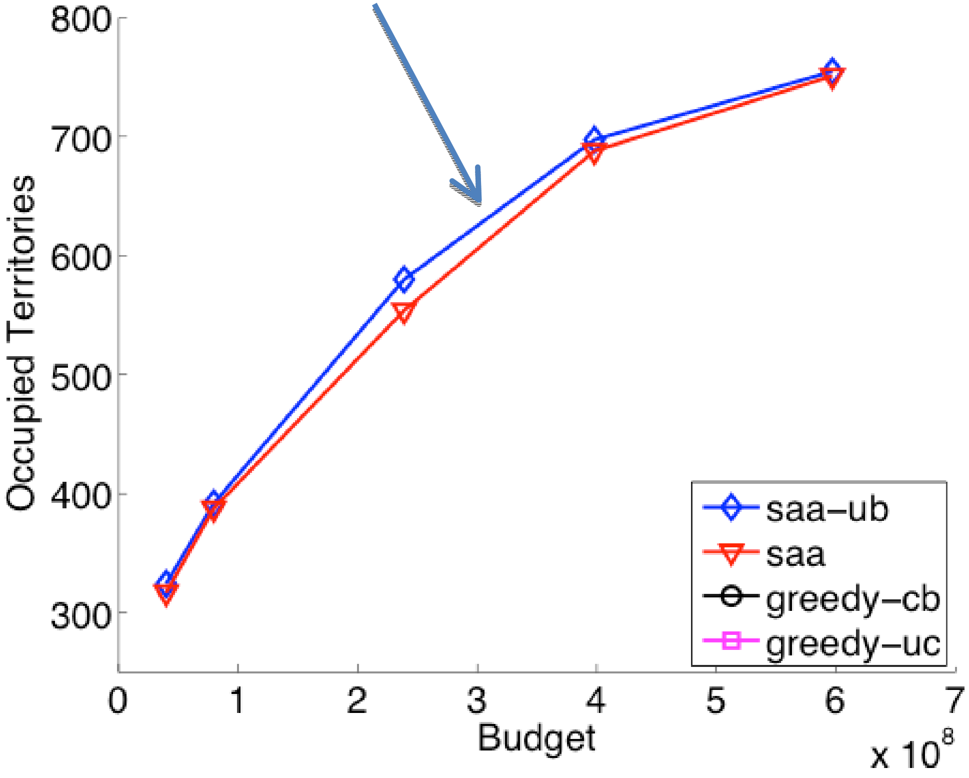

Upper bound!

$M = 50$, $N = 10$, $N_{test} = 500$

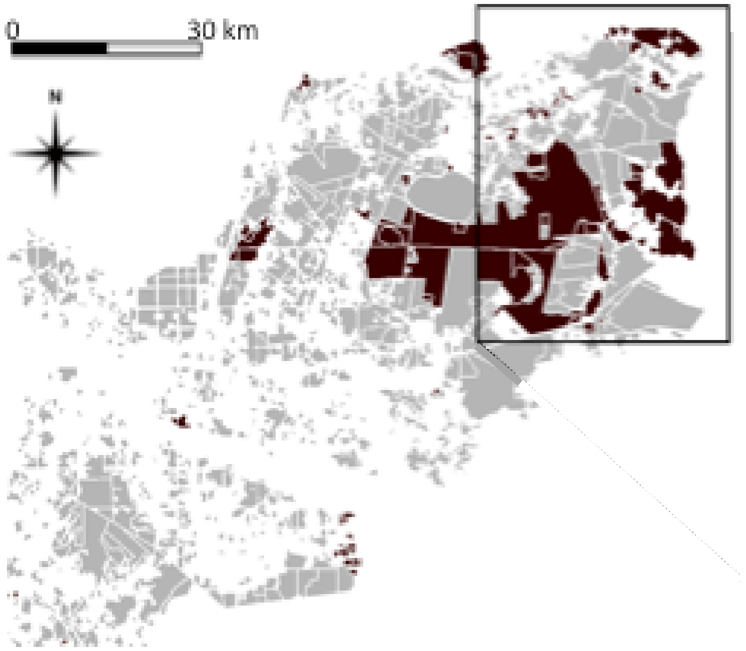


Results

Upper bound!

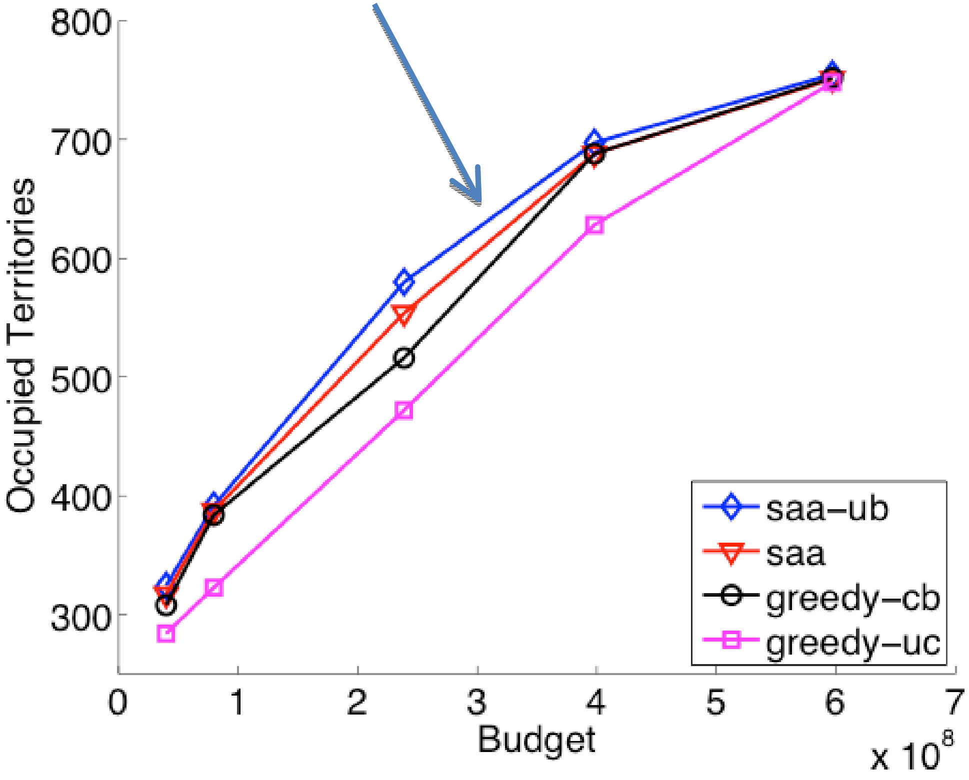


$M = 50$, $N = 10$, $N_{test} = 500$



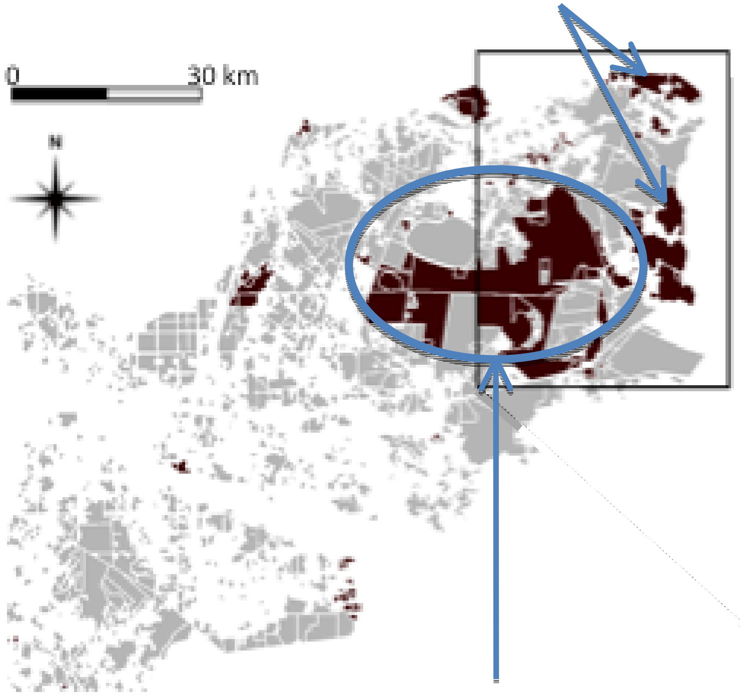
Results

Upper bound!



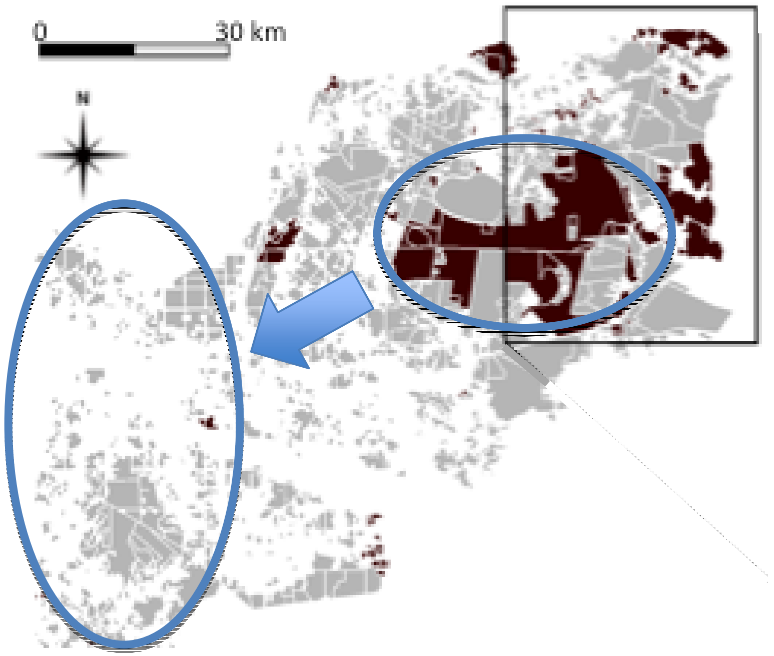
$M = 50$, $N = 10$, $N_{\text{test}} = 500$

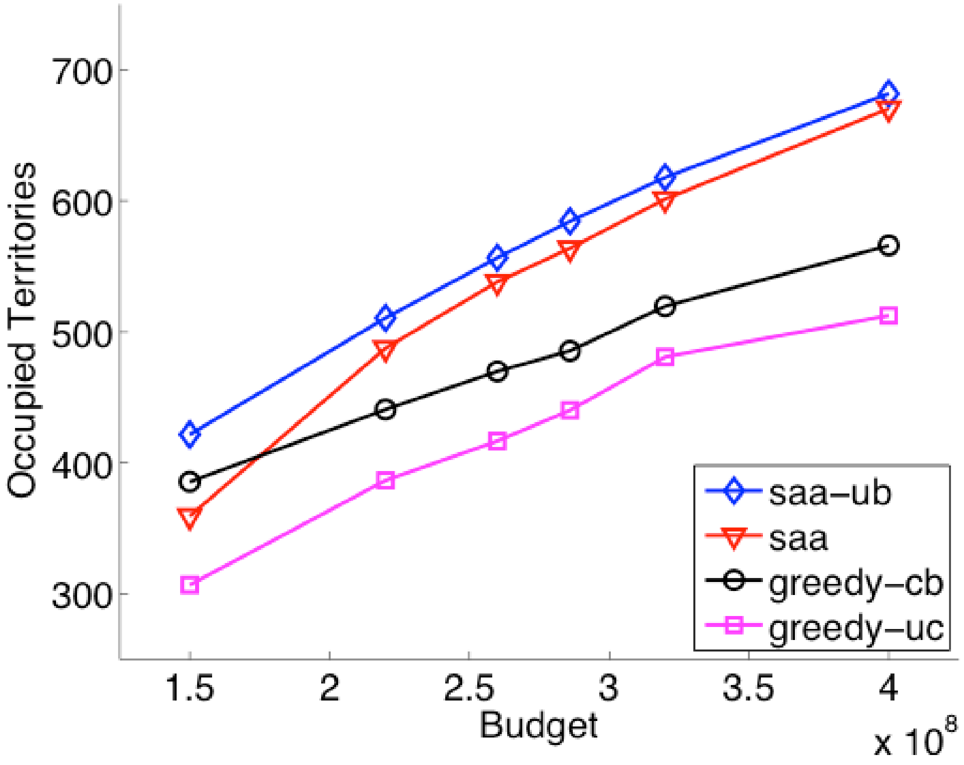
Initial population



Conservation Reservoir

A Harder Instance





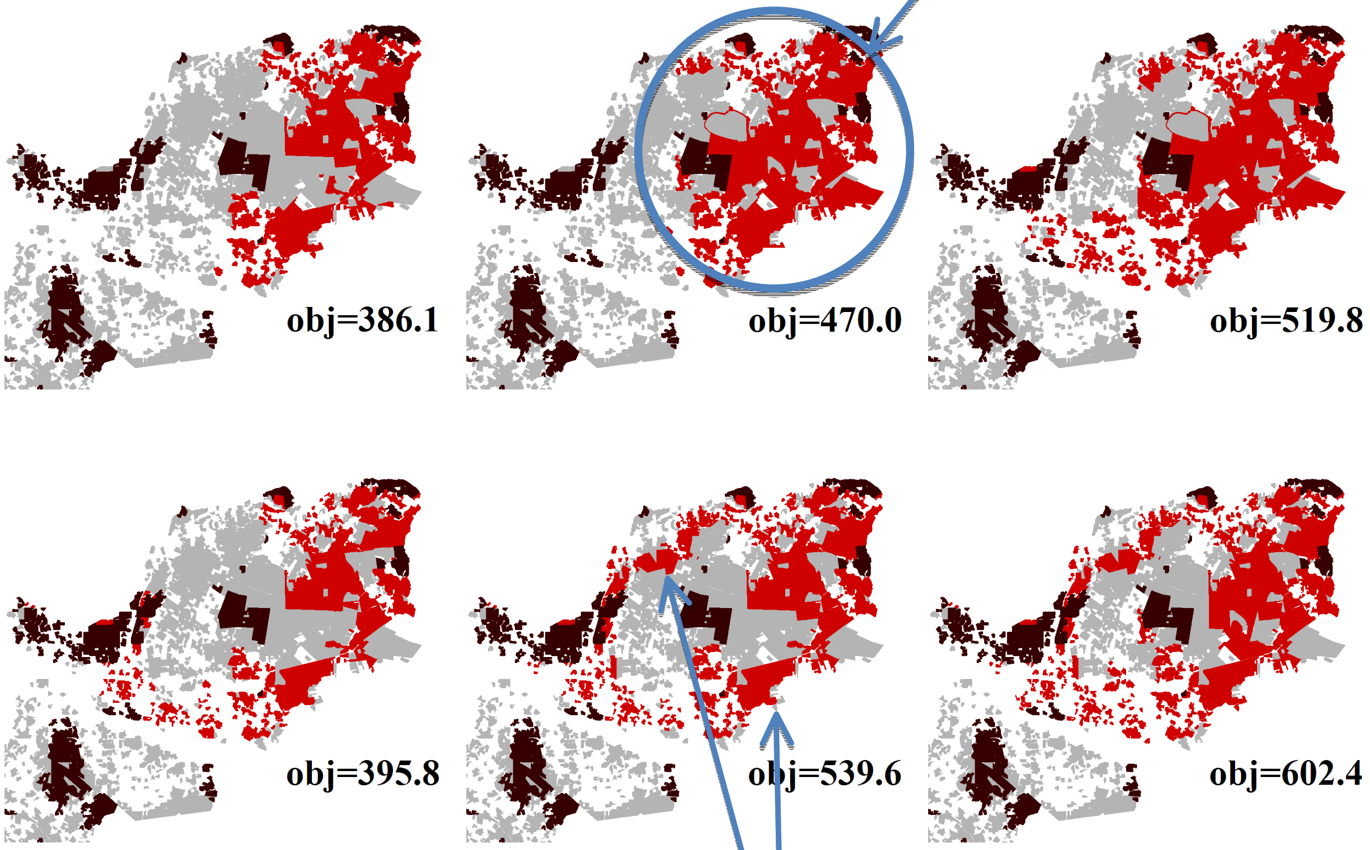
Move the conservation reservoir so it is more remote.

Conservation Strategies

Build outward from sources

Greedy Baseline

SAA Optimum (our approach)



$150M

$260M

$320M

Path-building (goal-setting)

Future Challenges

- The real world is complex
  - Multiple management agencies
  - Competing objectives
  - Budget comes in installments, not up front
  - Uncertainty about RCW dispersal behavior

- We want to provide computational tools that are useful in as many situations as possible; but a good model does **not** model everything

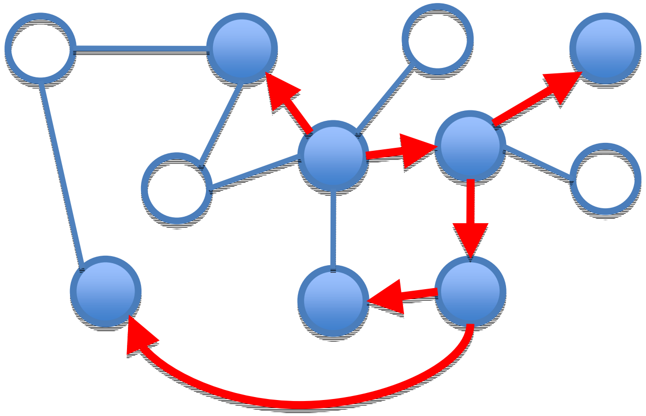
- What are the next steps?
  - Adaptive management
  - Solving the problem faster...

Thanks!

Previous Work: Viral Marketing

[Domingos and Richardson, KDD, 2001]

- Which customers to target to maximize the effectiveness of word-of-mouth marketing campaign?



Influence Maximization

[Kempe, Kleinberg, Tardos, KDD, 2003]

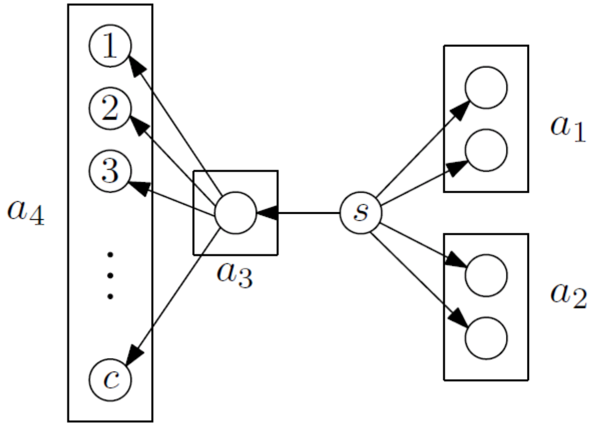
- Formalization of viral marketing as cascade optimization:
  - Choose *k* sources to maximize the spread of a cascade
  - (In our problem, equivalent to choosing initially occupied territories)
- Influence maximization is hard (NP-hard)!
  - But... it is easy to approximate, because it is *submodular*
  - Greedy algorithm!
- We are not so lucky
  - Objective is *not* submodular with respect to adding nodes
  - Greedy can do very poorly

Non-Submodularity

- Submodularity: diminishing returns property

$$S \subseteq T \quad \Rightarrow \quad f(S \cup \{m\}) - f(S) \geq f(T \cup \{m\}) - f(T).$$

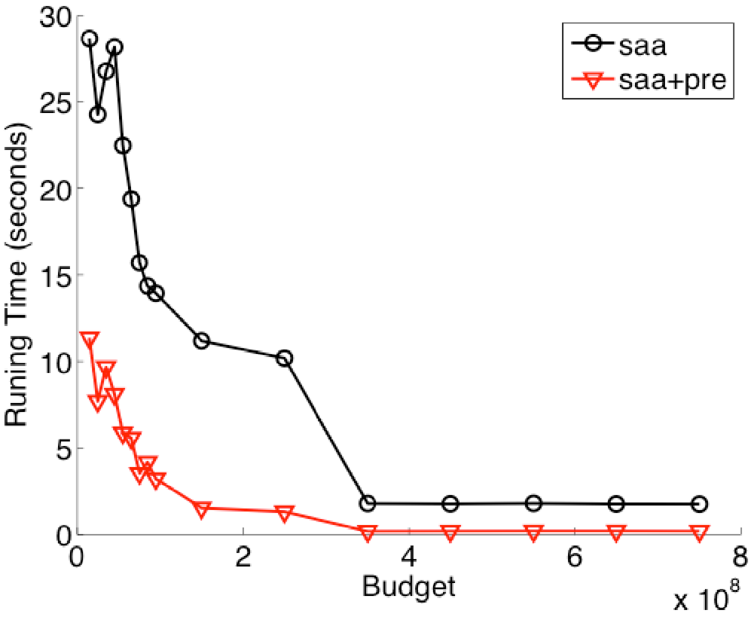
- Does not hold for actions that add nodes
  - Low payoff action enable high payoff action

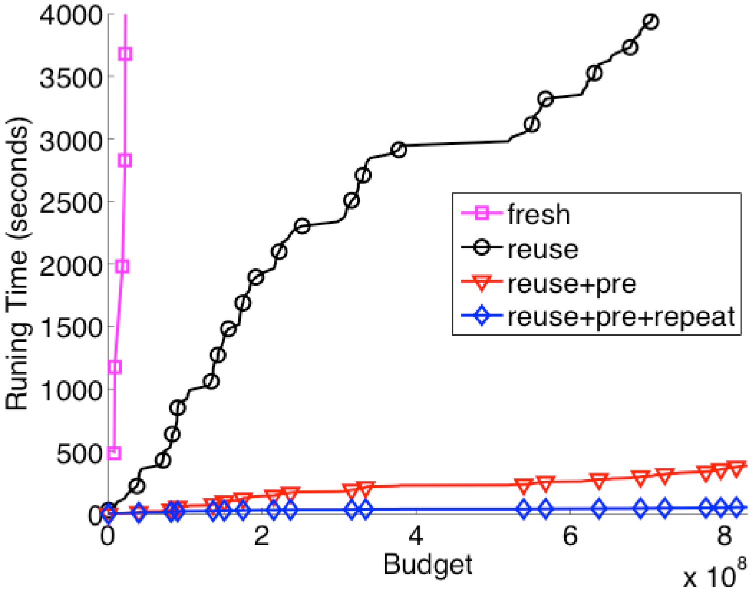


S	$f(S)$
$\{a_1, a_2\}$	4
$\{a_3\}$	1
$\{a_4\}$	0
$\{a_3, a_4\}$	$c + 1$

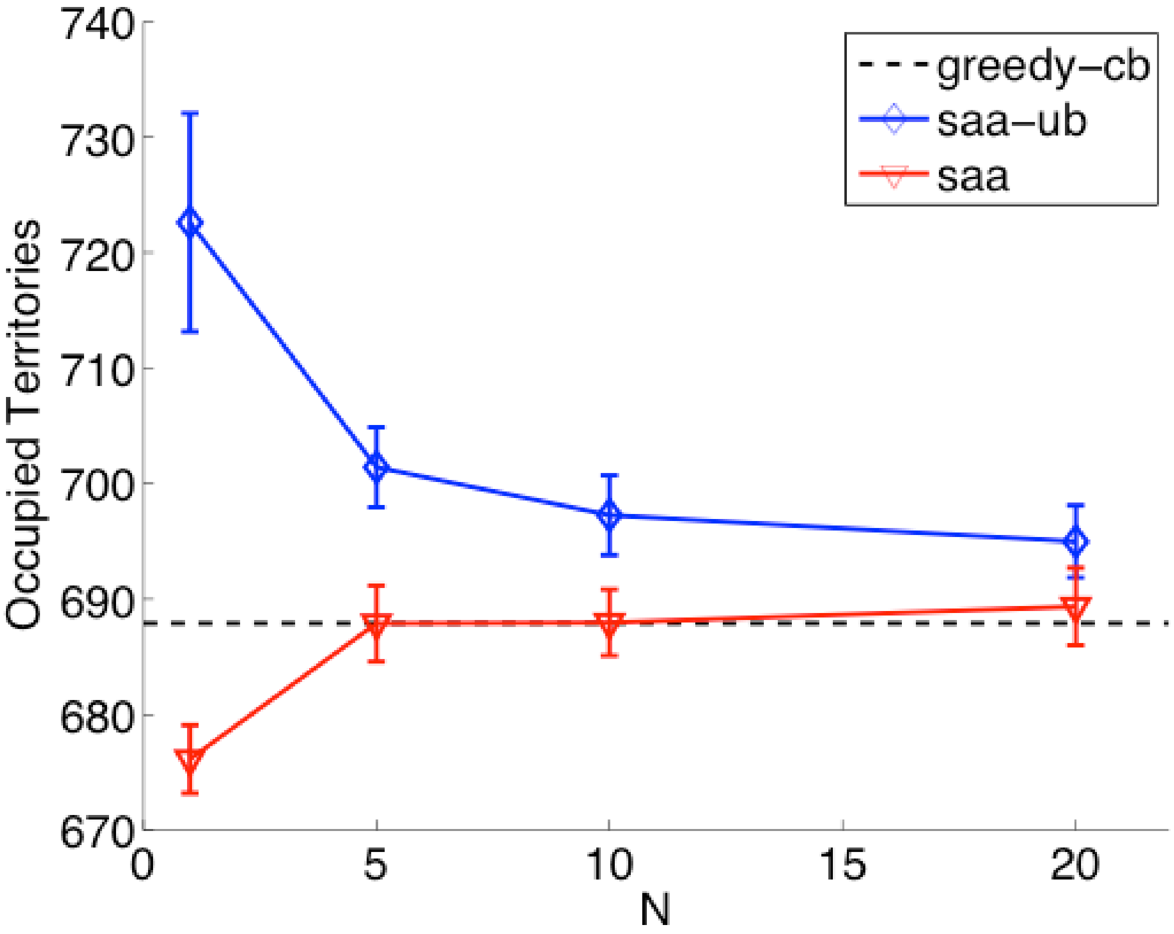
Preprocessing

- 2500 territories x 100 years x 10 training cascades = 2.5 milllion nodes
  - After pruning: 440K nodes
- Additional preprocessing steps reduce this to 200K nodes, 430K edges





SAA Convergence



Preprocessing

- Challenge: these are very big networks
- Create an equivalent but smaller network design problem
  1. Prune nodes that are not on any path from source to target
  2. Collapse nodes that are always reachable
  3. Collapse nodes whose “fates are tied” – reachable under identical sets of decisions