

LECTURE 10: INFLUENCE MAXIMIZATION IN NETWORKS

Prof. Pan Hui

CSIT 6000K: Social Networks and Social Computing: A Data Science Perspective

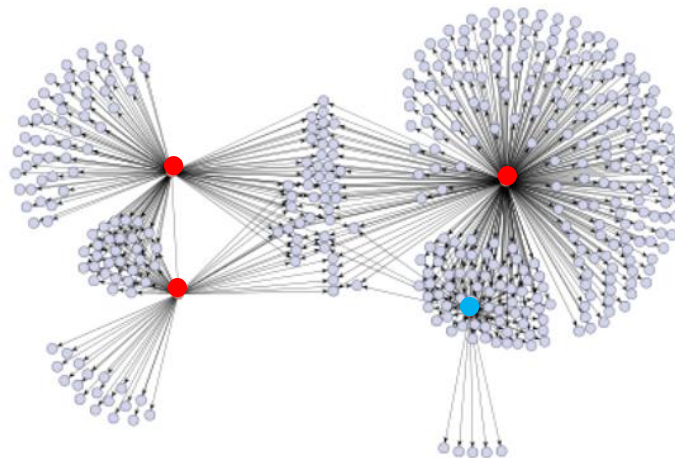
Thursdays 07:30 PM - 10:20 PM

How to Create Big Cascades?

2

□ Blogs – Information epidemics:

- Which are the influential blogs?
- Which blogs create big cascades?
- Where should we advertise?



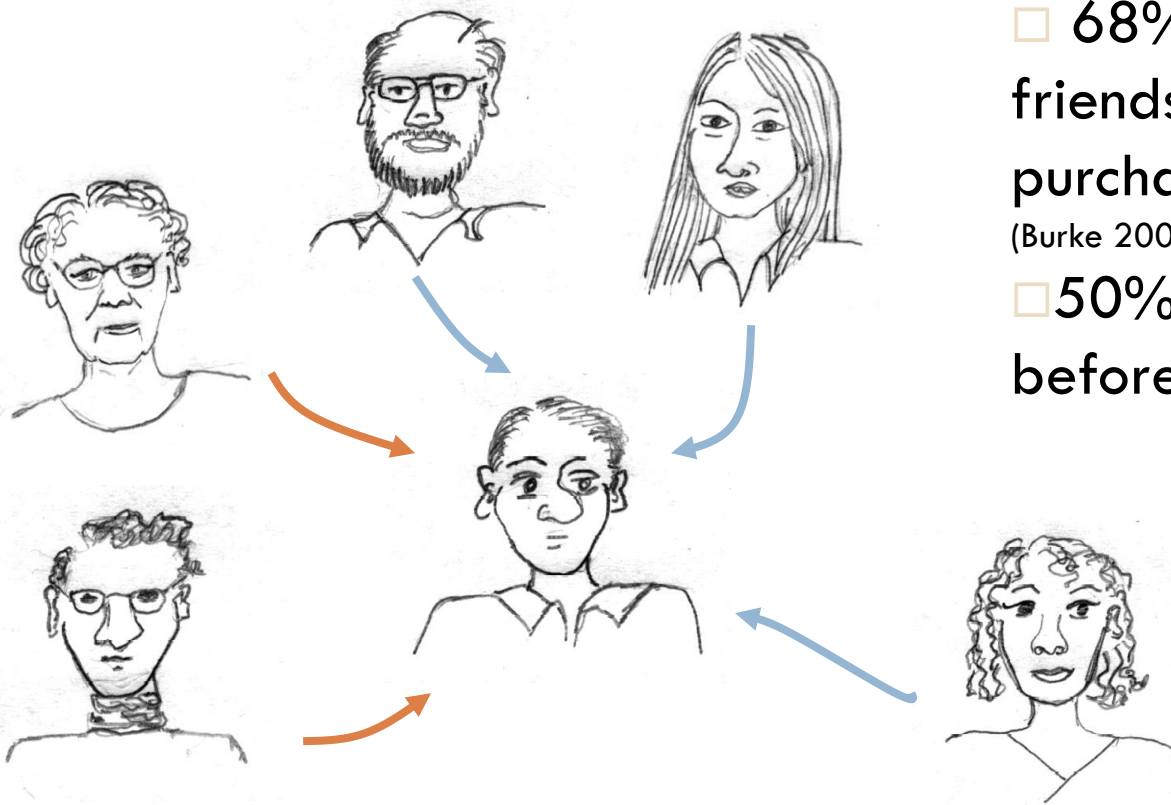
Which node shall we target?

● vs. ●

Viral Marketing?

3

- **We are more influenced by our friends than strangers**

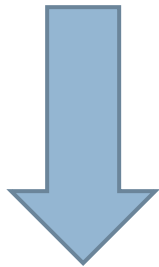


- 68% of consumers consult friends and family before purchasing home electronics (Burke 2003)
- 50% do research online before purchasing electronics

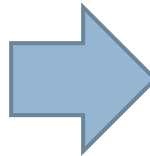
Viral Marketing

4

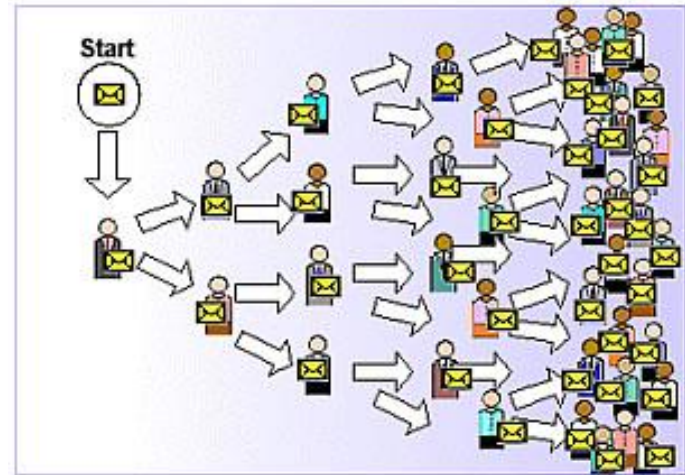
Identify influential customers



Convince them to adopt the product –
Offer discount/free samples



These customers endorse the product among their friends



Probabilistic Contagion

5

□ Independent Cascade Model

- Directed finite $G = (V, E)$
- Set S starts out with new behavior
 - Say nodes with this behavior are “active”
- Each edge (v, w) has a probability p_{vw}
- If node v is active, it gets one chance to make w active, with probability p_{vw}
 - Each edge fires at most once

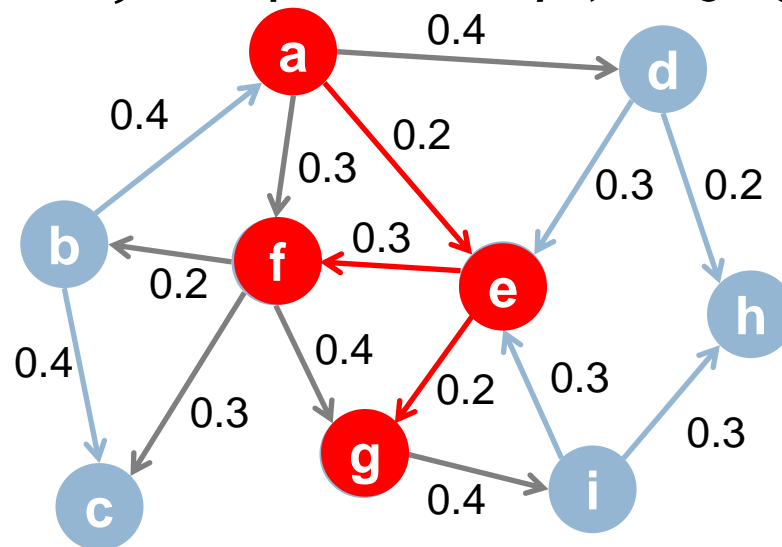
□ Does scheduling matter? **No**

- u, v both active, doesn't matter which fires first
- **But the time moves in discrete steps**

Independent Cascade Model

6

- Initially some nodes S are active
- Each edge (v, w) has probability (weight) p_{vw}

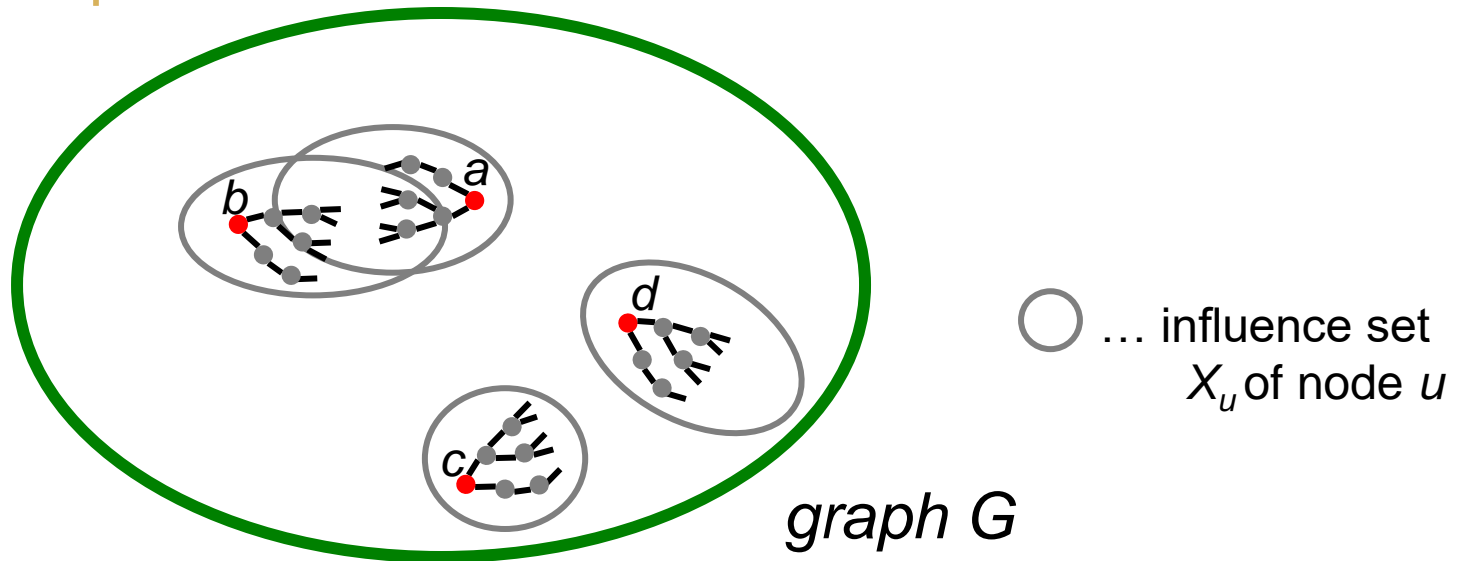


- When node v becomes active:
 - ▣ It activates each out-neighbor w with prob. p_{vw}
- Activations spread through the network

Most Influential Set of Nodes

7

- **S** : is initial active set
- **$f(S)$** : The expected size of final active set



- **Set S is more influential if $f(S)$ is larger**
$$f(\{a, b\}) < f(\{a, c\}) < f(\{a, d\})$$

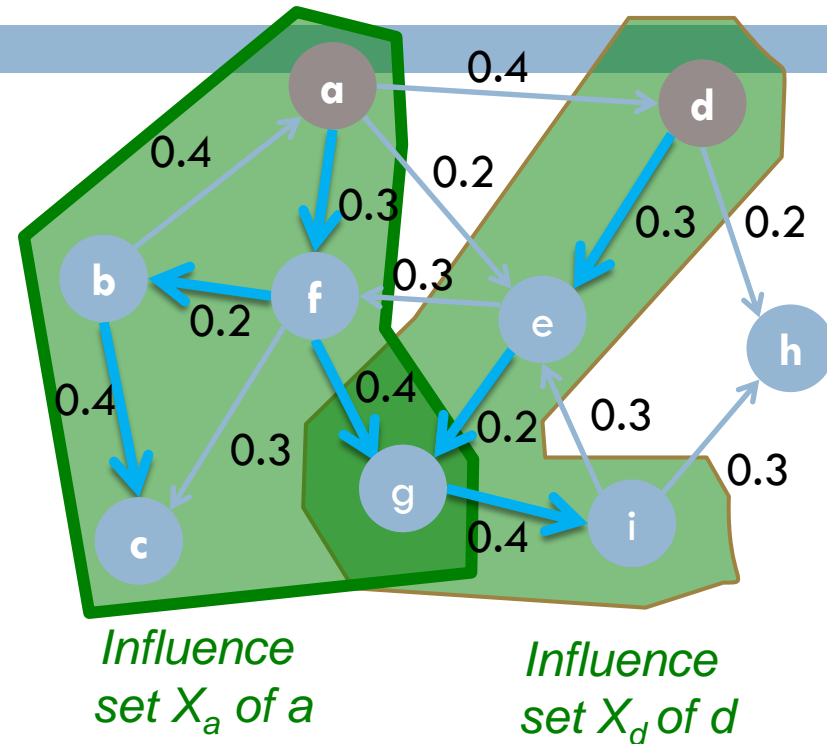
Most Influential Set

8

Emphasize that k is a parameter (given by the user)

Problem:

- **Most influential set of size k :** set S of k nodes producing **largest expected cascade size $f(S)$** if activated [Domingos-Richardson '01]
- **Optimization problem:**



$$\max_{S \text{ of size } k} f(S)$$

Why “expected cascade size”? X_a is a result of a random process. So in practice we would want to compute many realizations of X_a and then maximize the avg. $f(S)$

$$f(S) = \sum_{\text{Random realizations } i} f_i(S)$$

HOW HARD IS INFLUENCE
MAXIMIZATION?

Most Influential Subset of Nodes

10

- Most influential set of k nodes:
set S on k nodes producing largest expected cascade size $f(S)$ if activated
- **The optimization problem:**

$$\max_{S \text{ of size } k} f(S)$$

- **How hard is this problem?**
 - ▣ **NP-COMPLETE!**
 - Show that finding most influential set is at least as hard as a vertex cover

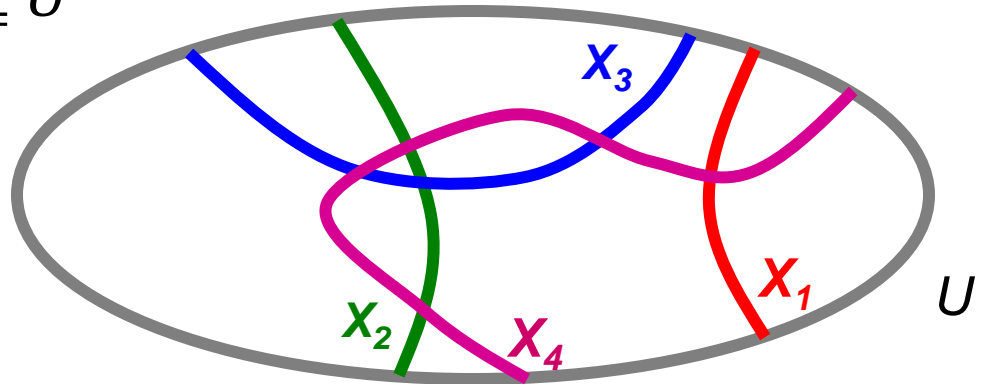
Background: Vertex Cover

11

□ Vertex cover problem

(a known NP-complete problem):

- Given universe of elements $U = \{u_1, \dots, u_n\}$ and sets $X_1, \dots, X_m \subseteq U$



- Are there k sets among X_1, \dots, X_m such that their union is U ?

□ Goal:

Encode vertex cover as an instance of

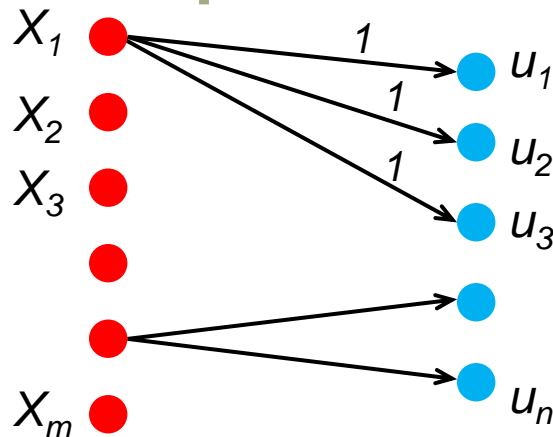
$$\max_{S \text{ of size } k} f(S)$$

Influence Maximization is NP-hard

12

□ Given a vertex cover instance with sets X_1, \dots, X_m

□ Build a bipartite “X-to-U” graph:



e.g.:
 $X_1 = \{u_1, u_2, u_3\}$

Construction:

- Create edge $(X_i, u) \forall X_i \forall u \in X_i$
-- directed edge from sets to their elements
- Put weight 1 on each edge (e.i., activation is deterministic)

□ **Vertex cover as Influence Maximization in X-to-U graph:** There exists a set S of size k with $f(S) = k + n$ iff there exists a size k set cover

Note: Optimal solution is always a set of sets X_i .

This problem is hard in general, could be special cases that are easier.

Summary so Far

13

- **Bad news:**
 - ▣ Influence maximization is NP-complete
- **Next, good news:**
 - ▣ There exists an approximation algorithm!
- **Consider the Hill Climbing algorithm to find S :**
 - ▣ **Input:**

Influence set of each node u : $X_u = \{v_1, v_2, \dots\}$

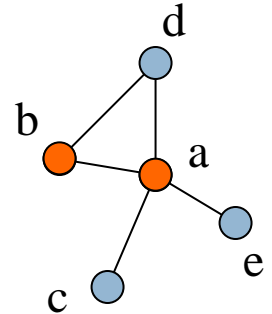
 - If we activate u , nodes $\{v_1, v_2, \dots\}$ will eventually get active
 - ▣ **Algorithm:** At each iteration i take the node u that gives best marginal gain: $\max_u f(S_{i-1} \cup \{u\})$
 - S_i ... Initially active set
 - $f(S_i)$... Size of the union of X_u , $u \in S_i$

(Greedy) Hill Climbing

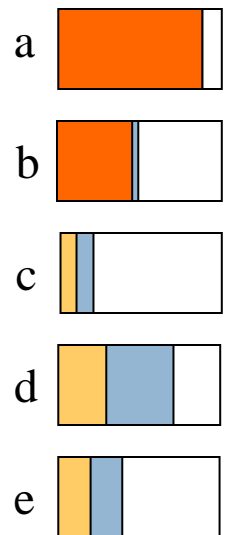
14

Algorithm:

- Start with $S_0 = \{ \}$
- For $i = 1 \dots k$
 - ▣ Take node u that $\max f(S_{i-1} \cup \{u\})$
 - ▣ Let $S_i = S_{i-1} \cup \{u\}$
- **Example:**
 - ▣ Eval. $f(\{a\}), \dots, f(\{e\})$, pick max of them
 - ▣ Eval. $f(\{a, b\}), \dots, f(\{a, e\})$, pick max
 - ▣ Eval. $f(\{a, b, c\}), \dots, f(\{a, b, e\})$, pick max



$f(S_{i-1} \cup \{u\})$



Approximation Guarantee

15

□ Hill climbing produces a solution S

where: $f(S) \geq (1 - 1/e) * \text{OPT}$ ($f(S) > 0.63 * \text{OPT}$)

[Nemhauser, Fisher, Wolsey '78, Kempe, Kleinberg, Tardos '03]

□ Claim holds for functions $f(\cdot)$ with 2 properties:

□ **f is monotone:** (activating more nodes doesn't hurt)

if $S \subseteq T$ then $f(S) \leq f(T)$ and $f(\{\}) = 0$

□ **f is submodular:** (activating each additional node helps less)

adding an element to a set gives less improvement
than adding it to one of its subsets: $\forall S \subseteq T$

$$\underbrace{f(S \cup \{u\}) - f(S)}_{\text{Gain of adding a node to a small set}} \geq \underbrace{f(T \cup \{u\}) - f(T)}_{\text{Gain of adding a node to a large set}}$$

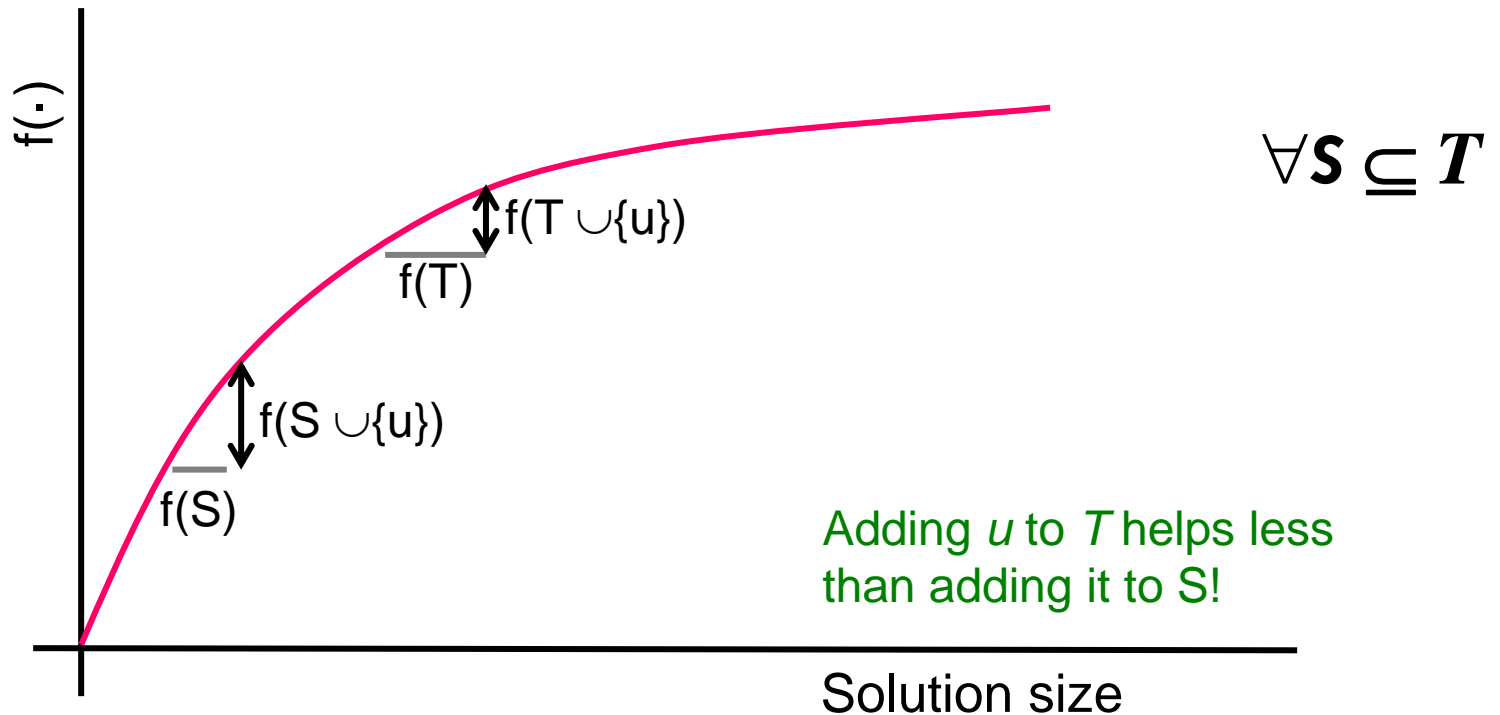
Gain of adding a node to a small set

Gain of adding a node to a large set

Submodularity– Diminishing returns

16

□ Diminishing returns:



$$\underbrace{f(S \cup \{u\}) - f(S)}_{\text{Gain of adding a node to a small set}} \geq \underbrace{f(T \cup \{u\}) - f(T)}_{\text{Gain of adding a node to a large set}}$$

Solution Quality

17

We just showed:

- Hill climbing finds solution S which
$$f(S) \geq (1 - 1/e) * \mathbf{OPT} \quad \text{i.e., } f(S) \geq 0.63 * \mathbf{OPT}$$
- This is a **data independent bound**
 - ▣ This is a worst case bound
 - ▣ No matter what is the input data (influence sets), we know that the Hill-Climbing won't never do worse than $0.63 * \mathbf{OPT}$

HOMOPHILY AND SOCIAL INFLUENCE



Birds of a Feather

20

- Similarity breeds connection. This principle—the homophily principle—structures network ties of every type, including marriage, friendship, work, advice, support, information transfer, exchange, comembership, and other types of relationship. The result is that people's personal networks are homogeneous with regard to many sociodemographic, behavioral, and intrapersonal characteristics. Homophily limits people's social worlds in a way that has powerful implications for the information they receive, the attitudes they form, and the interactions they experience. Homophily in race and ethnicity creates the strongest divides in our personal environments, with age, religion, education, occupation, and gender following in roughly that order. Geographic propinquity, families, organizations, and isomorphic positions in social systems all create contexts in which homophilous relations form. Ties between nonsimilar individuals also dissolve at a higher rate, which sets the stage for the formation of niches (localized positions) within social space.

Homophily

21

- Agents in a social network have **other characteristics** apart from their links
 - ▣ Non-mutable: race, gender, age
 - ▣ Mutable: place to live, occupation, activities, opinions, beliefs
- Links and mutable characteristics co-evolve over time

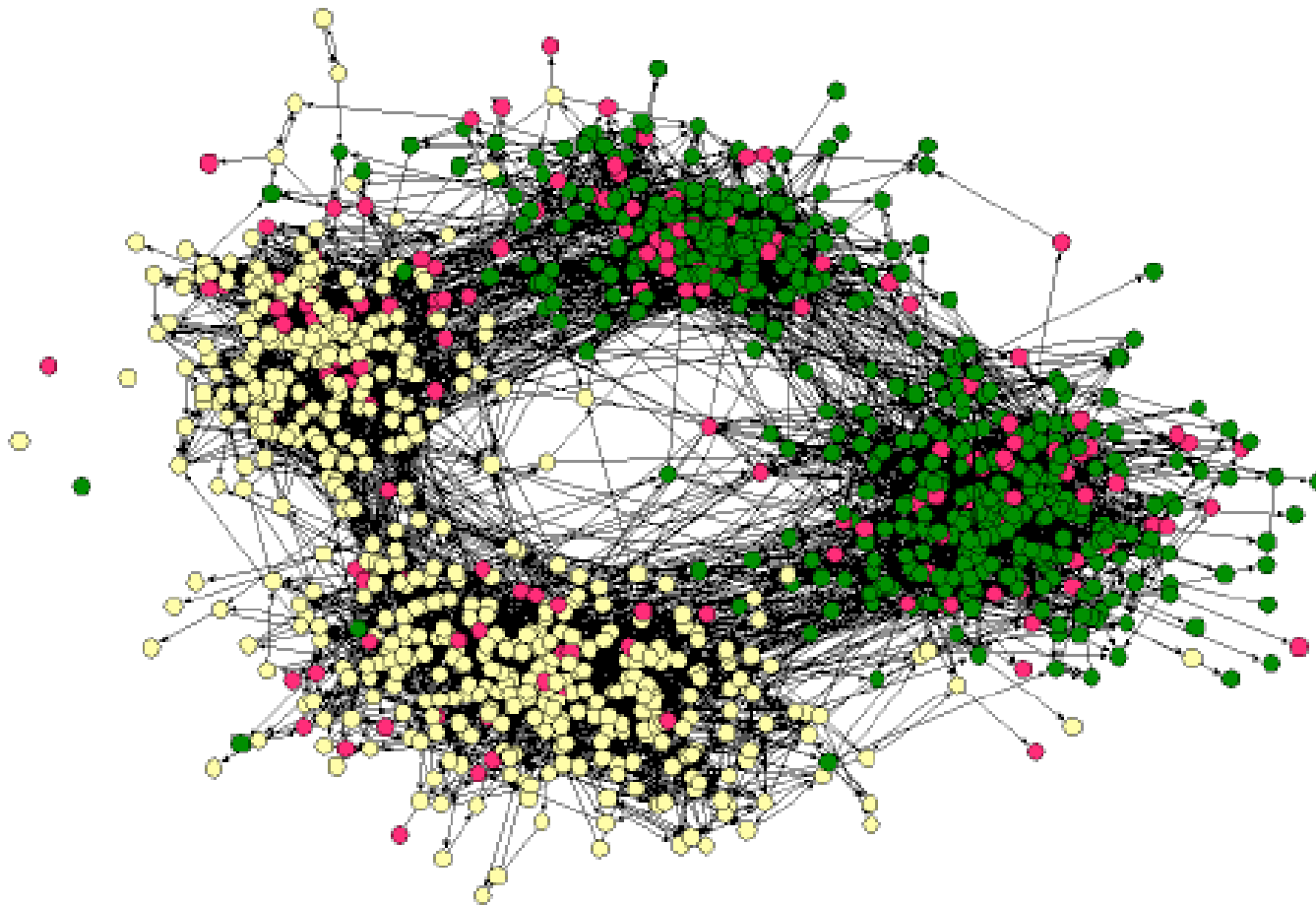
Homophily

22

- When we take a snapshot in time, we observe that these node characteristics are correlated across links
 - ▣ E.g. Academics have often academic friends, etc.
- This phenomenon that people are linked to similar others is called **homophily**

Homophily at a U.S. High School

23



Homophily

24

- Mechanisms underlying Homophily
 - ▣ Selection
 - A and B have similar characteristics -> A and B form a link AB
 - ▣ Social Influence
 - A and B have a link -> B chooses the same (mutable) characteristic as A
 - E.g. A starts smoking, and B follows (peer pressure)

Social-Affiliation Network

25

- Network of persons and social *foci* (activities)

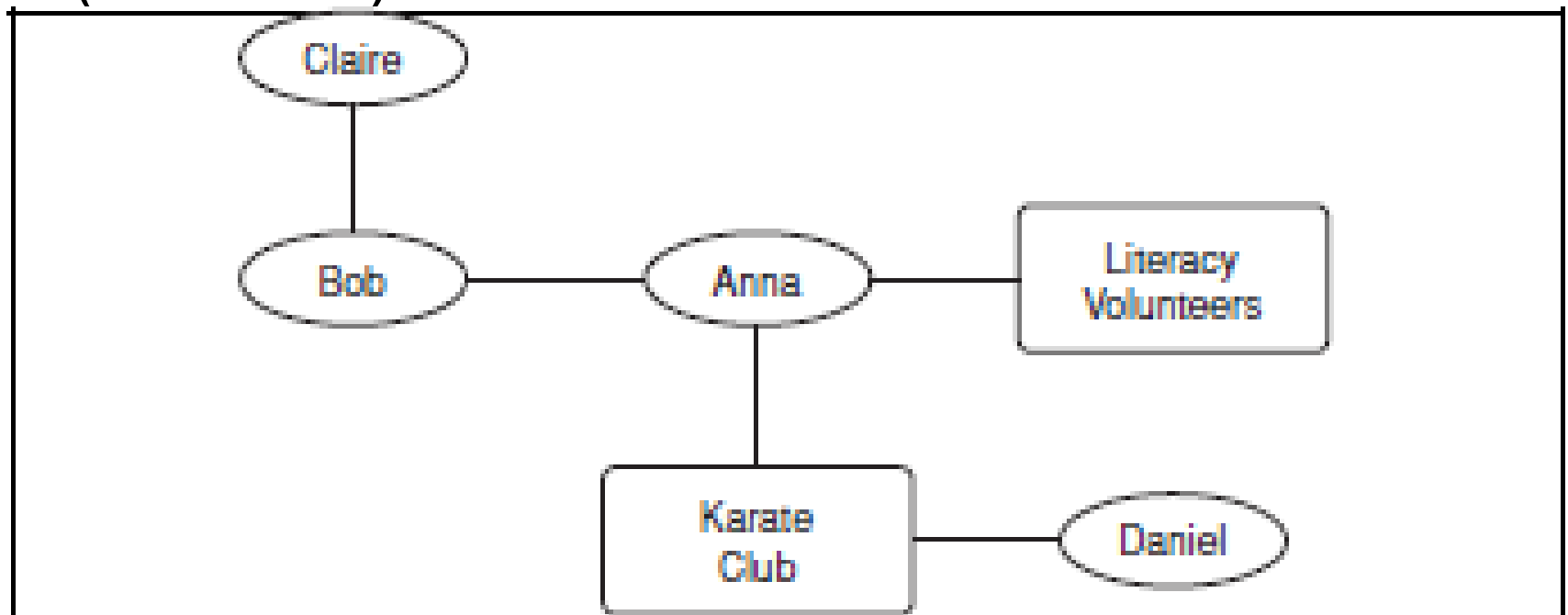


Figure 4.5. A social-affiliation network shows both the friendships between people and their affiliation with different social foci.

Triadic Closure

26

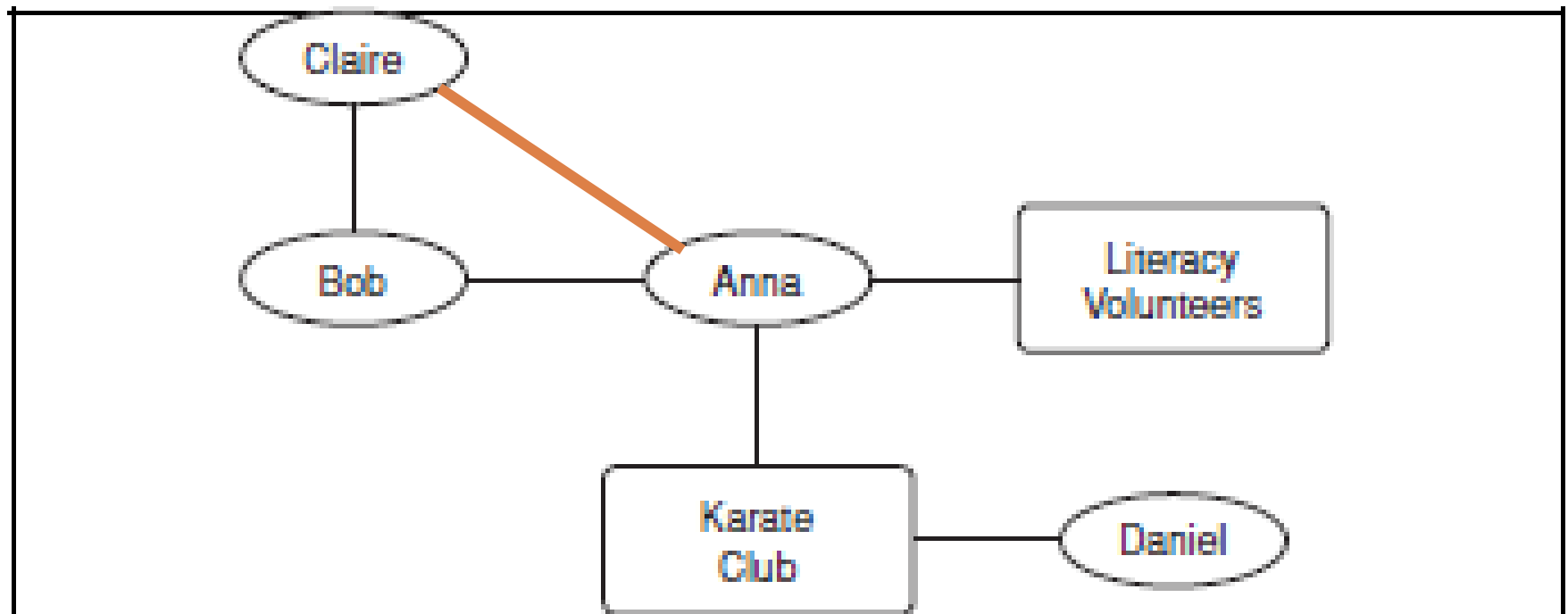


Figure 4.5. A social-affiliation network shows both the friendships between people and their affiliation with different social foci.

Focal Closure

27

- **Selection:** *Karate* introduces *Anna* to *Daniel*

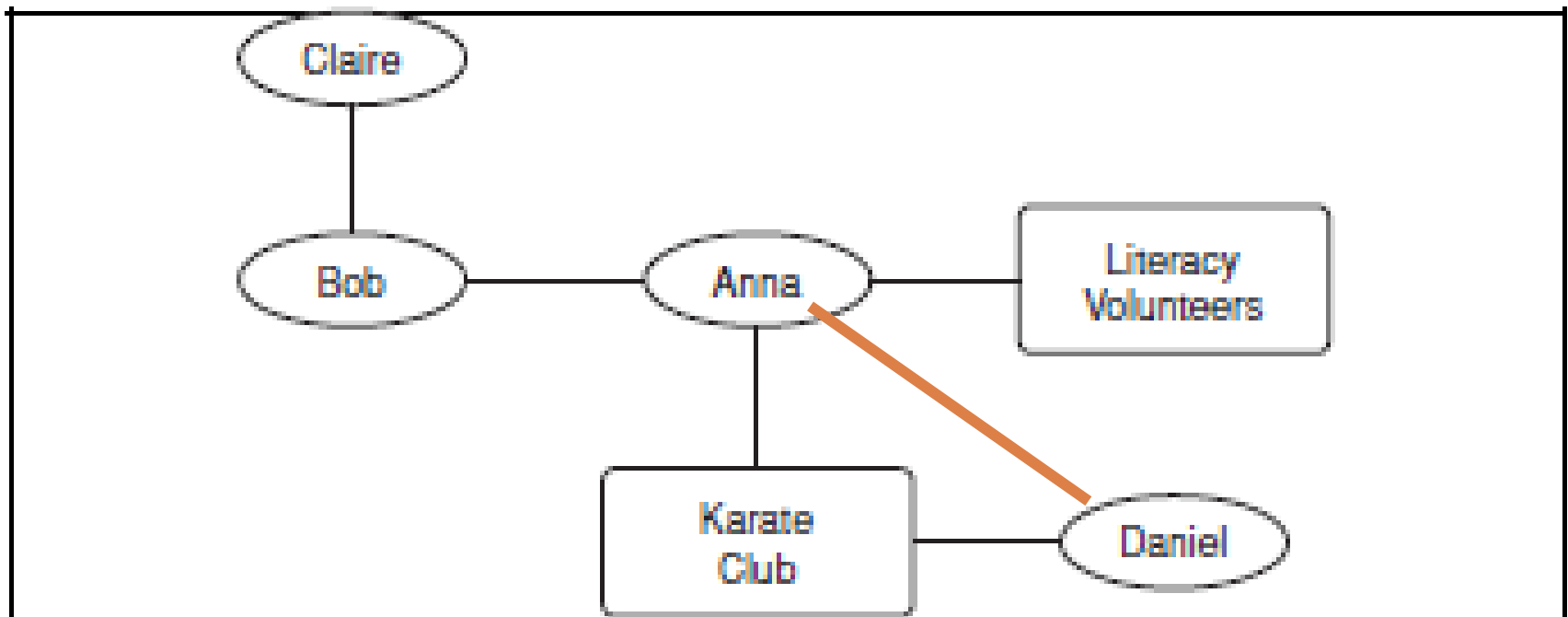


Figure 4.5. A social-affiliation network shows both the friendships between people and their affiliation with different social foci.

Membership Closure

28

- **Social Influence:** *Anna* introduces *Bob* to *Karate*

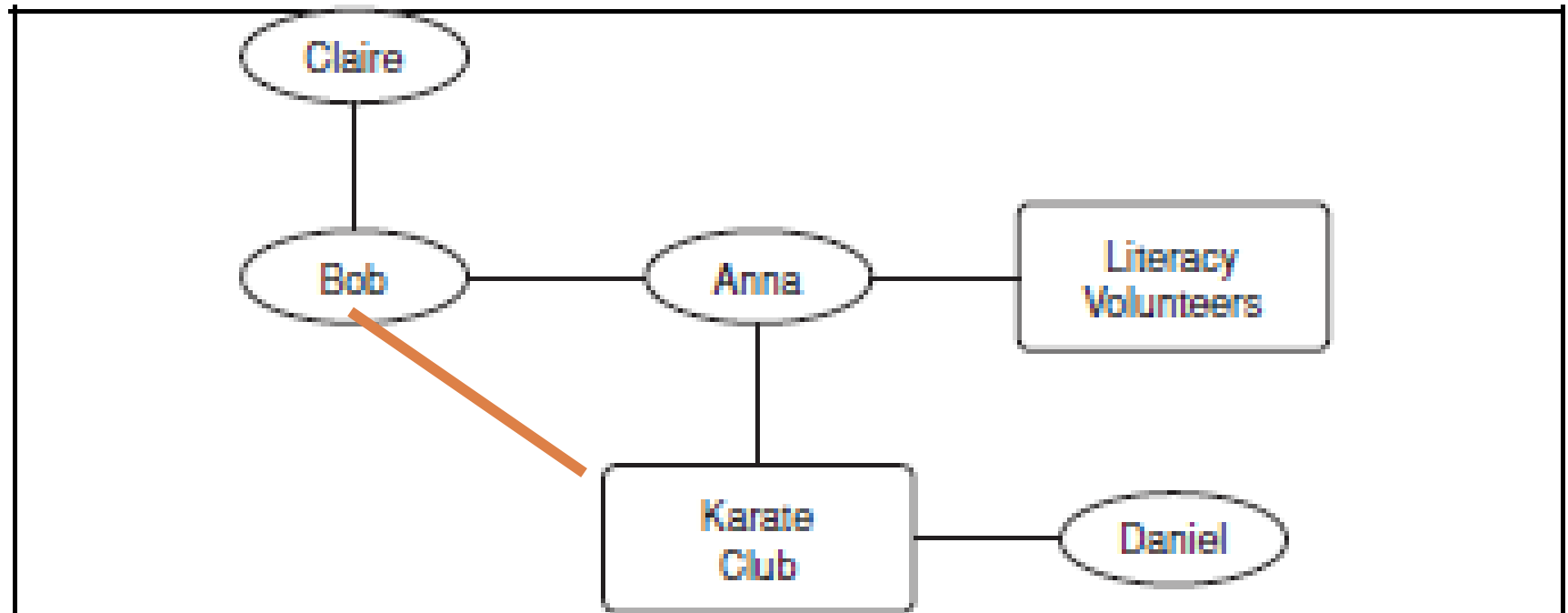


Figure 4.5. A social-affiliation network shows both the friendships between people and their affiliation with different social foci.

Homophily

29

- Both Selection and Social Influence drive homophily
- How important is each mechanism?
 - ▣ Important question: Different mechanism implies different policy,
 - e.g. Policy to prevent teenagers from smoking
 - Social Influence. Target “key players” and let them positively influence rest
 - Selection. Target on characteristics (e.g. family background) alone

Homophily

30

- Both Selection and Social Influence drive homophily
- How important is each mechanism?
 - ▣ Difficult question:
 - Requires longitudinal data
 - Requires observation of (almost) all characteristics
 - If a characteristic is not observed, then social influence effect is overestimated

Homophily

31

- Measuring the mechanisms behind homophily is a hot topic
 - Kossinets & Watts (2006): Detailed course and e-mail interaction data from university
 - Centola (2010, 2011): Experimental data on social influence controlling network structure
 - Sacerdote (2001): Social influence among students after randomized dorm assignment

Homophily and Segregation

33

- Neighborhoods tend to be segregated according to race or culture
 - ▣ Ghetto formation
 - ▣ What is the mechanism behind that?

Segregation in Chicago

33



(a) *Chicago, 1940*



(b) *Chicago, 1960*

Homophily and Segregation

34

- Segregation model of Thomas Schelling
 - ▣ Agent-based model
 - Two different agents: X and O types
 - Agents live on a grid
 - **weak satisficing preferences for homophily**
 - At least k of the 8 neighbors of same type
 - Each period, agents who are not satisfied move to a location where they are
























Schelling's model ($k=3$)

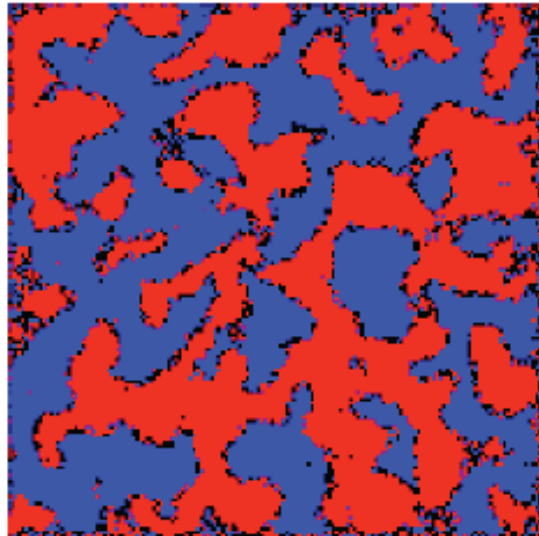
35

X	X				
X	O		O		
X	X	O	O	O	
X	O			X	X
	O	O	X	X	X
		O	O	O	

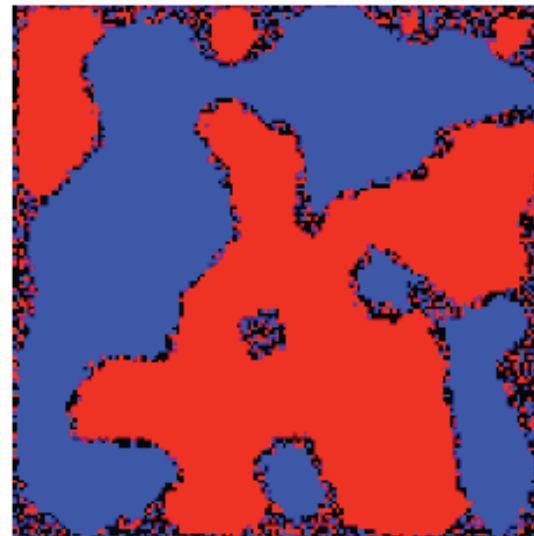
Schelling's model ($k=3$)

36

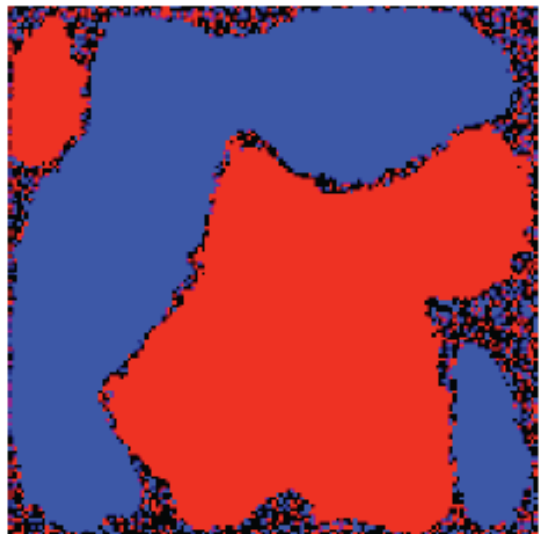
					
					
					
					
					
					



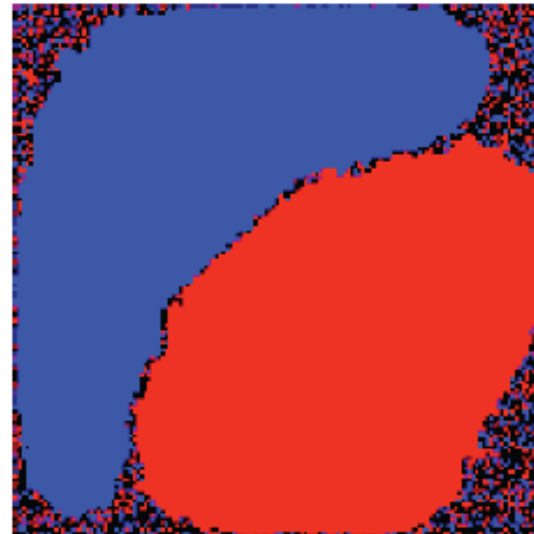
(a) After 20 steps



(b) After 150 steps



(c) After 350 steps



(d) After 800 steps

Schelling's model

38

- Surprising relation between **micro-behavior** and **macro-outcomes**
 - ▣ Weak satisficing preferences for homophily sufficient to create complete segregation
 - ▣ Segregation arises due to miscoordination
 - There exists an allocation involving **complete integration satisfying all agents**, but individual decisionmaking does not lead to that outcome