Segregation & Link Prediction

Advanced Social Computing

Department of Computer Science
University of Massachusetts, Lowell
Spring 2020

Hadi Amiri
hadi@cs.uml.edu

Lecture Topics

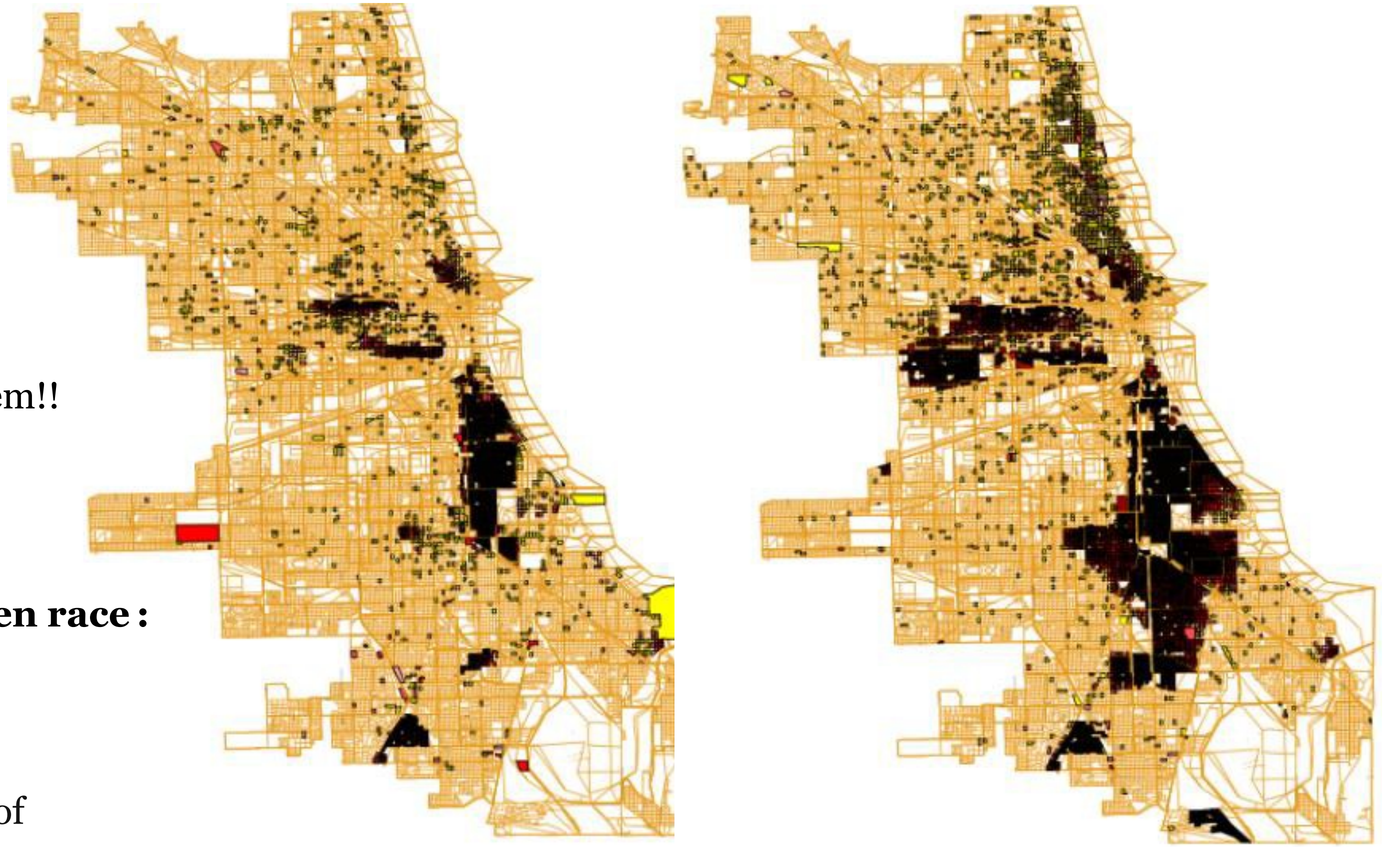
- Spatial Model of Segregation
- Link Prediction in Social Networks

Spatial Model of Segregation

Effects of homophily
in the formation of ethnically
and racially **homogeneous**
neighborhoods in cities.

People live near others like them!!

Color the map wrt to a given race :
--Lighter: Lowest percentage
of the race
--Darker: highest percentage of
the race.



(a) *Chicago, 1940*

(b) *Chicago, 1960*

Schelling Model

- How **global patterns** of spatial segregation arise from the effect of homophily operating at a **local level**.
  - Many factors affect
    - Focus on an intentionally simplified mechanism
  - Forces leading to segregation are robust!
    - Operate even when no one individual explicitly wants a segregated outcome!

Schelling Model- Basics

- Let's assume a population of individuals called agents
  - agents of type **X** or **O**
- The two types represent some immutable characteristic as the basis for homophily
  - race, ethnicity, country of origin, or native language

x	x				
x	o		o		
x	x	o	o	o	
x	o			x	x
	o	o	x	x	x
		o	o	o	

(a) *Agents occupying cells on a grid.*

Schelling Model- Basics- Cnt.

- Agents reside in cells of a grid
  - 2-dimentional geography of a city
- Some cells are unpopulated
- Cell's neighbors: cells that touch it including diagonal contact
  - cells not on boundary: 8 neighbors

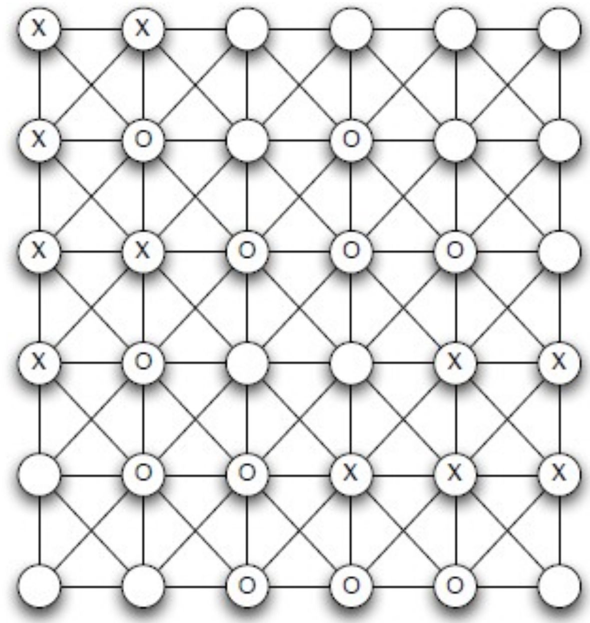
X	X				
X	O		O		
X	X	O	O	O	
X	O			X	X
	O	O	X	X	X
		O	O	O	

(a) *Agents occupying cells on a grid.*

Schelling Model- Basics- Cnt.

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

(a) *Agents occupying cells on a grid.*



(b) *Neighbor relations as a graph.*

- Create a network by:
  - considering cells as the nodes, and
  - putting an edge between two cells that are neighbors on the grid!

Schelling Model- Constraints

- The fundamental constraint driving the model:
  - Each agent wants to have at least *t* other agents of its own type as neighbors.
  - Otherwise, it will be unsatisfied
    - Move to a new location that makes it satisfied!

Schelling Model- Constraints

- t=3
  - Unsatisfied nodes *

X1*	X2*				
X3	O1*		O2		
X4	X5	O3	O4	O5*	
X6*	O6			X7	X8
	O7	O8	X9*	X10	X11
		O9	O10	O11*	

Schelling Model- Movements

- Unsatisfied agents move in rounds
  - Considered in some order
  - Move to unoccupied cells where they become satisfied!
    - Cells that satisfies them:
      - a random cell, or the nearest cell, or sweep downward along rows, etc.

Schelling Model- Movements- Cnt.

- Moves may make other agents unsatisfied
  - Leads to a new round of movement:
    - Other agents move to become satisfied!
  - Deadlocks may happen
    - Agent need to move but there is no cell to make it satisfied:
      - Stay where it is, or moved to a completely random cell!

Schelling Model- Constraints

X1*	X2*				
X3	O1*		O2		
X4	X5	O3	O4	O5*	
X6*	O6			X7	X8
	O7	O8	X9*	X10	X11
		O9	O10	O11*	

- t=3, Unsatisfied *
- Order:
  - one row at a time working downward!
- Moves:
  - nearest cell!

X3	X6	O1	O2		
X4	X5	O3	O4		
	O6	X2	X1	X7	X8
O11	O7	O8	X9	X10	X11
	O5	O9	O10*		

Schelling Model- Constraints

X1*	X2*				
X3	O1*		O2		
X4	X5	O3	O4	O5*	
X6*	O6			X7	X8
	O7	O8	X9*	X10	X11
		O9	O10	O11*	

- t=3, Unsatisfied *
- Order:
  - one row at a time working downward!
- Moves:
  - nearest cell!

- Are agents more segregated now?

X3	X6	O1	O2		
X4	X5	O3	O4		
	O6	X2	X1	X7	X8
O11	O7	O8	X9	X10	X11
	O5	O9	O10*		

Schelling Model- Constraints

X1*	X2*				
X3	O1*		O2		
X4	X5	O3	O4	O5*	
X6*	O6			X7	X8
	O7	O8	X9*	X10	X11
		O9	O10	O11*	

- t=3, Unsatisfied *
- Order:
  - one row at a time working downward!
- Moves:
  - nearest cell!

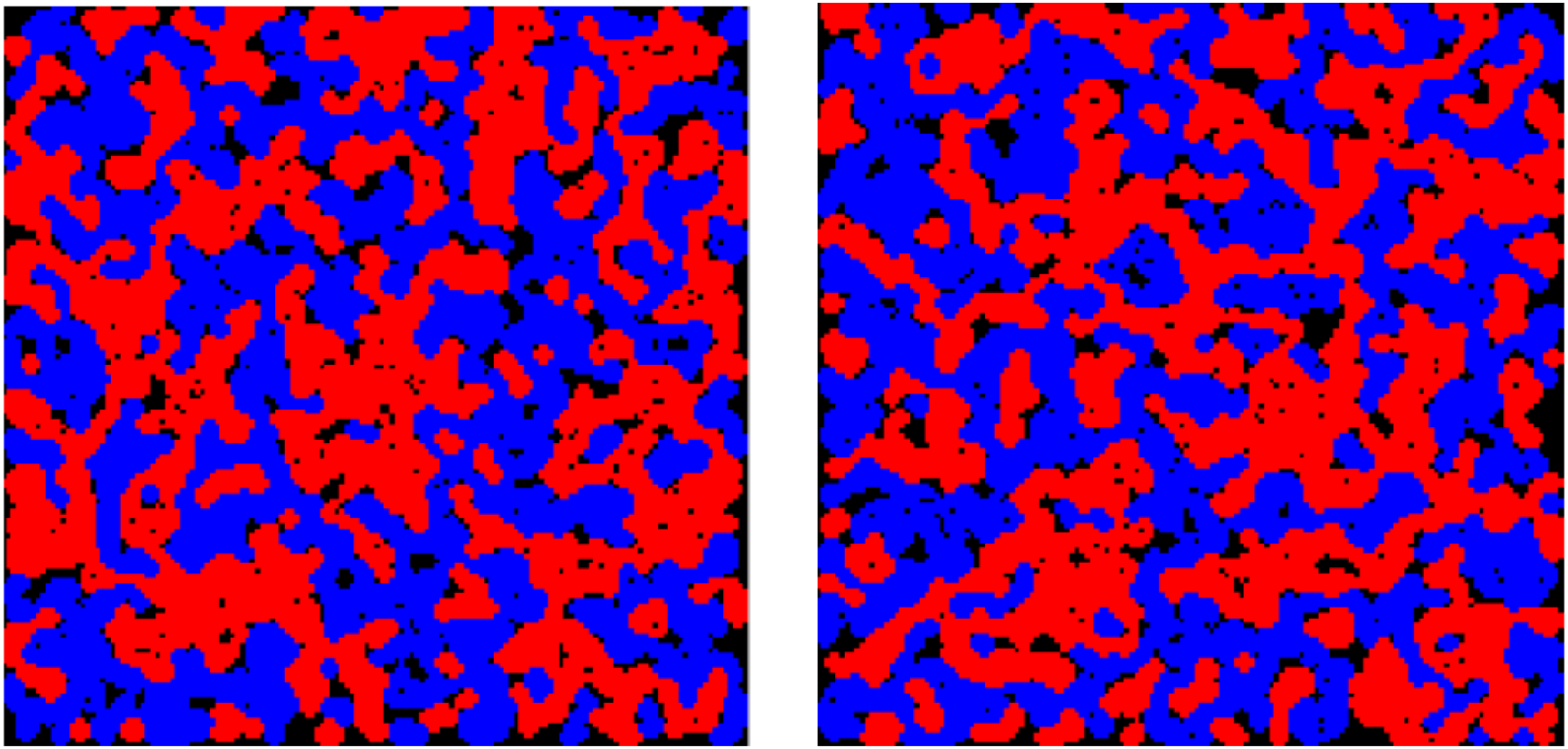
- Are agents more segregated now?
  - (a) 1 agent with no neighbors of the opposite type
  - (b) six such agents

X3	X6	O1	O2		
X4	X5	O3	O4		
	O6	X2	X1	X7	X8
O11	O7	O8	X9	X10	X11
	O5	O9	O10*		

Schelling Model- Movements

- Qualitative results of the model tend to be quite similar!

Schelling Model- Simulation 1

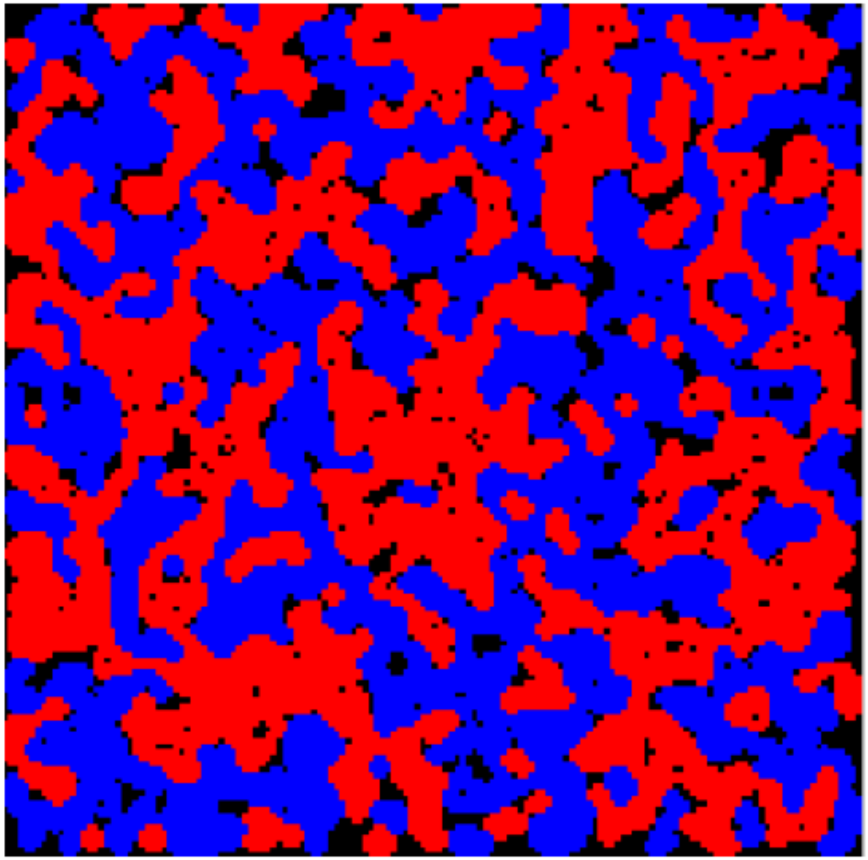


(a) *A simulation with threshold 3.* (b) *Another simulation with threshold 3.*

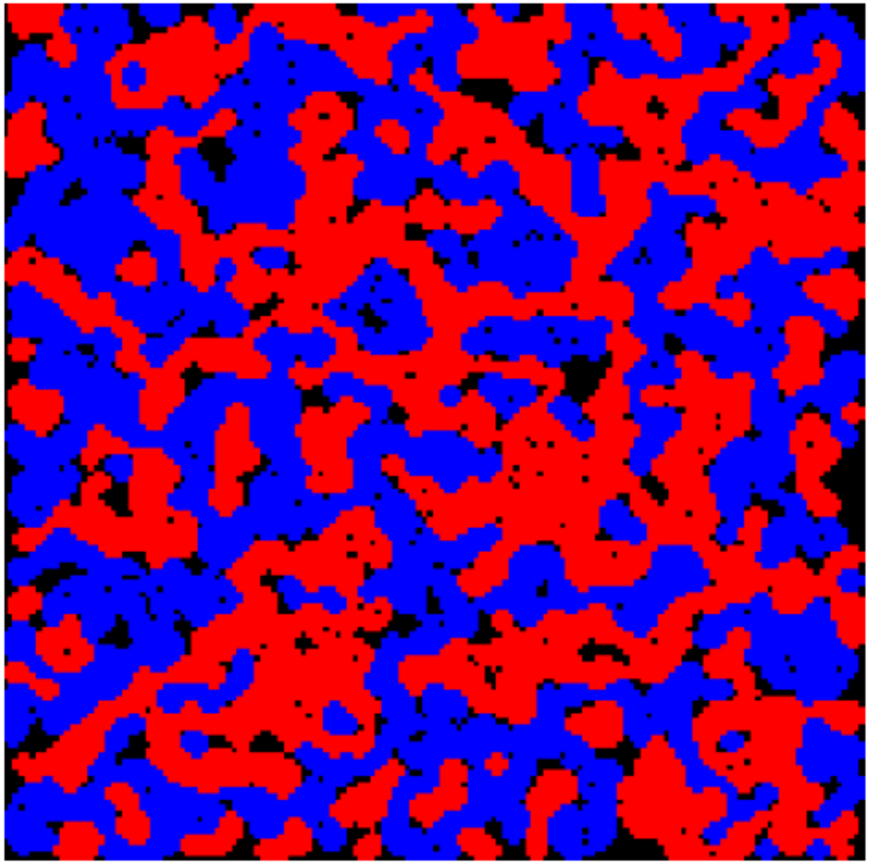
t=3, 150-by-150 grid with 10,000 agents of each type. Random Starting pattern!

X: Red **O: Blue** **Not occupied: Black**

Schelling Model- Simulation 1- Cnt.



(a) *A simulation with threshold 3.*



(b) *Another simulation with threshold 3.*

Large homogeneous regions interlocking with each other!
Large numbers of agents surrounded by agents of same type!

Schelling Model- Interpretation

- Segregation is taking place even though no individual agent is seeking it:
  - agents just want to be near *t* others like them
  - t=3 → agents are willing to be in the minority
    - 3 neighbors of its own type, 5 neighbors of opposite type

- Segregation is not happening because we have subtly built into the model!

Schelling Model- Interpretation- Cnt.

- A checkerboard 4*4 pattern
  - all agent are satisfied
  - agents not on the boundary have exactly 4 neighbors of each type.
- Why don't we observe these kinds of patterns in simulations?

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

Schelling Model- Interpretation- Cnt.

- Why don't we observe these kinds of patterns in simulations?
  - It is hard to find such integrated patterns from a **random start**.
  - Agents attach themselves to **clusters** of others like themselves (**higher probability** to be satisfied).
  - Agent movements cause previously satisfied agents to fall below the threshold and move as well (**Progressive unraveling**).

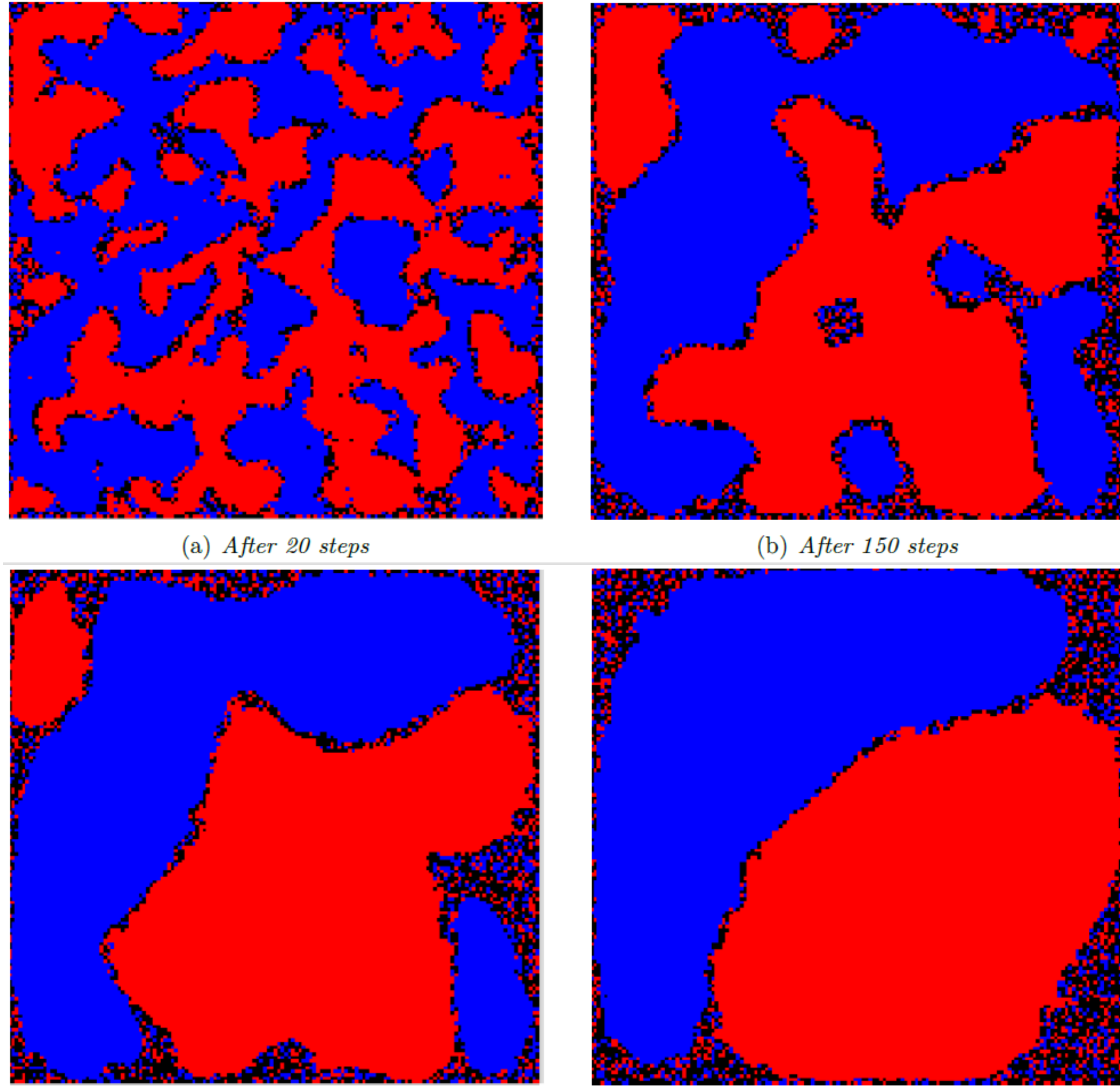
Simulation 2

- t=4

Nodes are willing to have equal number of neighbors of each type!

- iterations:
    - 20
    - 150
    - 350
    - 800!

Four intermediate points in one run of a simulation



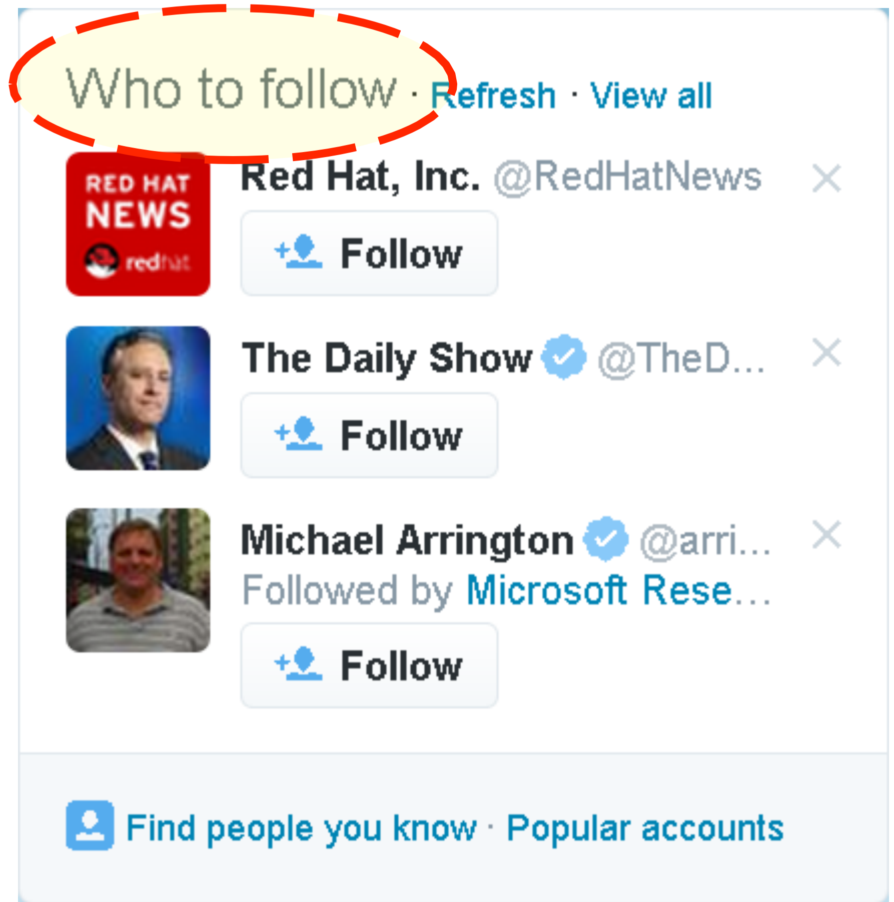
(a) *After 20 steps* (b) *After 150 steps* (c) *After 350 steps* (d) *After 800 steps*

Any integration among the two types tends to collapse completely over time.

Schelling Model- Interpretation- Cnt.

- The overall effect:
  - **Local Preferences** of individual agents have produced a **Global Pattern** that none of them necessarily intended.
  - Immutable characteristics can become highly correlated with mutable characteristics (here decision about where to live).

Link Prediction



Link Prediction- Problem

- Link prediction problem
  - Given a snapshot of a network, infer which new links between nodes are likely to occur in the future?

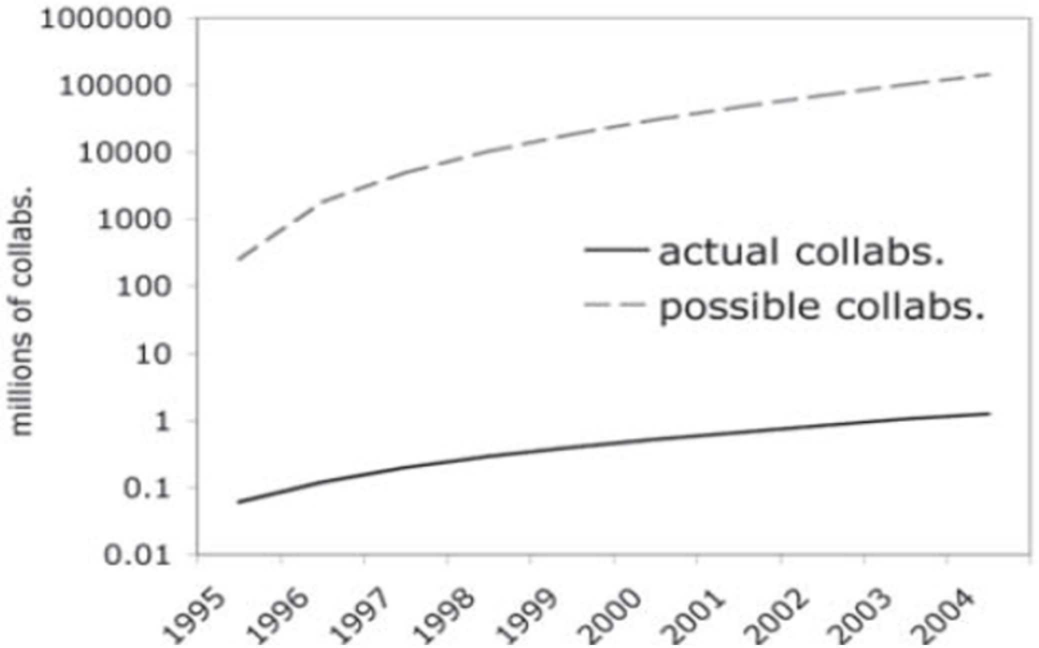
- To what extent link formation can be modeled using features that are **intrinsic** to the network itself?

- Compute **proximity of nodes** in a network.

The link-prediction problem for social networks. Liben-Nowell et. al. J of American society for info science and technology. 2007.

Link Prediction – Challenges

- Large class skewness
  - #of possible edges is quadratic in the #of nodes, but only a tiny fraction of these edges are added to the graph!



- Nature of Collab?
- Author increase through time
- Richer experience through time

Figure 9.1. Log plot of actual and possible collaborations between DBLP authors, 1995-2004.

Source: Rattigan, M., et al. The case for anomalous link discovery. SIGKDD'05.

Link Prediction – Challenges – Cnt.

- Model calibration
  - The process of finding the function that transforms the output score value of the model to a label.
  - Sometimes more crucial than finding a good model.
    - False negatives are catastrophic in detecting links in a terrorist network.
    - False positives are worse than false negative in recommending friendship links.

- Training cost in terms of time complexity

- Need for dynamic updating of model

Link Prediction- Algorithm

Algorithm

\---------------

1. Take the input graph → training data
2. Pick a pair of nodes (x, y)
3. Assign link btw x and y a weight: *score*(x, y)
4. Develop supervised classifiers
   - Develop features
   - Make a list in descending order of *score*(.,.) values!
5. Evaluate with test graph (data)

How to compute *score*(.,.)?

Measures of Proximity

graph distance	(negated) length of shortest path between x and y
common neighbors	$\lvert\Gamma(x) \cap \Gamma(y)\rvert$
Jaccard's coefficient	$\frac{\lvert\Gamma(x)\cap\Gamma(y)\rvert}{\lvert\Gamma(x)\cup\Gamma(y)\rvert}$
Adamic/Adar	$\sum_{z\in\Gamma(x)\cap\Gamma(y)} \frac{1}{\log\lvert\Gamma(z)\rvert}$
preferential attachment	$\lvert\Gamma(x)\rvert \cdot \lvert\Gamma(y)\rvert$
Katz$_\beta$	$\sum_{\ell=1}^{\infty} \beta^\ell \cdot \lvert\mathsf{paths}_{x,y}^{\langle\ell\rangle}\rvert$ where $\mathsf{paths}_{x,y}^{\langle\ell\rangle}$:= {paths of length exactly ℓ from x to y} weighted: $\mathsf{paths}_{x,y}^{\langle 1\rangle}$:= number of collaborations between x, y. unweighted: $\mathsf{paths}_{x,y}^{\langle 1\rangle}$:= 1 iff x and y collaborate.
hitting time	$-H_{x,y}$
stationary-normed	$-H_{x,y} \cdot \pi_y$
commute time	$-(H_{x,y} + H_{y,x})$
stationary-normed	$-(H_{x,y} \cdot \pi_y + H_{y,x} \cdot \pi_x)$ where $H_{x,y}$:= expected time for random walk from x to reach y π_y := stationary distribution weight of y (proportion of time the random walk is at node y)
rooted PageRank$_\alpha$	stationary distribution weight of y under the following random walk: with probability α, jump to x. with probability $1-\alpha$, go to random neighbor of current node.
SimRank$_\gamma$	$\begin{cases} 1 & \text{if } x = y \\ \gamma \cdot \frac{\sum_{a\in\Gamma(x)}\sum_{b\in\Gamma(y)} \mathsf{score}(a,b)}{\lvert\Gamma(x)\rvert\cdot\lvert\Gamma(y)\rvert} & \text{otherwise} \end{cases}$

Link Prediction- Data

	training period			Core		
	authors	papers	edges	authors	$\lvert E_{old} \rvert$	$\lvert E_{new} \rvert$
`astro-ph`	5343	5816	41852	1561	6178	5751
`cond-mat`	5469	6700	19881	1253	1899	1150
`gr-qc`	2122	3287	5724	486	519	400
`hep-ph`	5414	10254	17806	1790	6654	3294
`hep-th`	5241	9498	15842	1438	2311	1576

Figure 1: The five sections of the arXiv from which co-authorship networks were constructed: **astro-ph** (astrophysics), **cond-mat** (condensed matter), **gr-qc** (general relativity and quantum cosmology), **hep-ph** (high energy physics—phenomenology), and **hep-th** (high energy physics—theory). The set Core is the subset of the authors who have written at least $\kappa_{training} = 3$ papers during the training period and $\kappa_{test} = 3$ papers during the test period. The sets E_{old} and E_{new} denote edges between Core authors which first appear during the training and test periods, respectively.

Link Prediction- Performance

predictor		astro-ph	cond-mat	gr-qc	hep-ph	hep-th
probability that a random prediction is correct		0.475%	0.147%	0.341%	0.207%	0.153%
graph distance (all distance-two pairs)		*9.6*	*25.3*	*21.4*	*12.2*	*29.2*
common neighbors		**18.0**	**41.1**	**27.2**	**27.0**	**47.2**
preferential attachment		4.7	6.1	7.6	*15.2*	7.5
Adamic/Adar		*16.8*	**54.8**	**30.1**	**33.3**	**50.5**
Jaccard		*16.4*	**42.3**	19.9	**27.7**	*41.7*
SimRank	$\gamma = 0.8$	*14.6*	*39.3*	*22.8*	*26.1*	*41.7*
hitting time		6.5	23.8	*25.0*	3.8	13.4
hitting time—normed by stationary distribution		5.3	23.8	11.0	11.3	21.3
commute time		5.2	15.5	**33.1**	*17.1*	23.4
commute time—normed by stationary distribution		5.3	16.1	11.0	11.3	16.3
rooted PageRank	$\alpha = 0.01$	*10.8*	*28.0*	**33.1**	*18.7*	*29.2*
	$\alpha = 0.05$	*13.8*	*39.9*	**35.3**	*24.6*	*41.3*
	$\alpha = 0.15$	*16.6*	**41.1**	**27.2**	**27.6**	*42.6*
	$\alpha = 0.30$	*17.1*	**42.3**	*25.0*	**29.9**	*46.8*
	$\alpha = 0.50$	*16.8*	**41.1**	*24.3*	**30.7**	*46.8*
Katz (weighted)	$\beta = 0.05$	3.0	21.4	19.9	2.4	12.9
	$\beta = 0.005$	*13.4*	**54.8**	**30.1**	*24.0*	**52.2**
	$\beta = 0.0005$	*14.5*	**54.2**	**30.1**	**32.6**	**51.8**
Katz (unweighted)	$\beta = 0.05$	*10.9*	**41.7**	**37.5**	*18.7*	**48.0**
	$\beta = 0.005$	*16.8*	**41.7**	**37.5**	*24.2*	**49.7**
	$\beta = 0.0005$	*16.8*	**41.7**	**37.5**	*24.9*	**49.7**

The link-prediction problem for social networks. Liben-Nowell et. al. J of American society for info science and technology. 2007.

Link Prediction - Twitter

- Read this paper/watch the talk on Twitter's practical approach to link prediction!
  - Gupta, P., et al. Wtf: The who to follow service at twitter. WWW'13.
  - https://www.youtube.com/watch?v=ZvXDLhqFkhc

Link Prediction - Counteracting

- Private connections can be exposed by LP algs and individuals can mitigate such threats.
- How can individuals rewire their connections to hide their sensitive relationships?

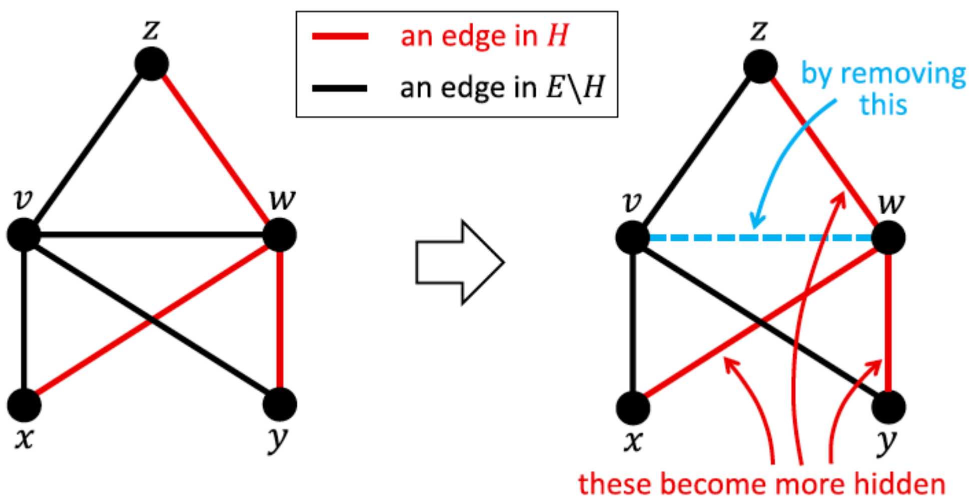


Figure 1. An illustration of the main idea behind the CTR heuristic. Here, by removing (v, w), we remove from the network three closed triads: one containing the nodes v, w, x, another containing v, w, y, and a third containing v, w, z. Consequently, the similarity scores of (x, w), (w, y) and (w, z) can only decrease based on the analysis in *Materials and Methods*.

Link Prediction - Counteracting

- Private connections can be exposed by LP algs and individuals can mitigate such threats.
- How can individuals rewire their connections to hide their sensitive relationships?

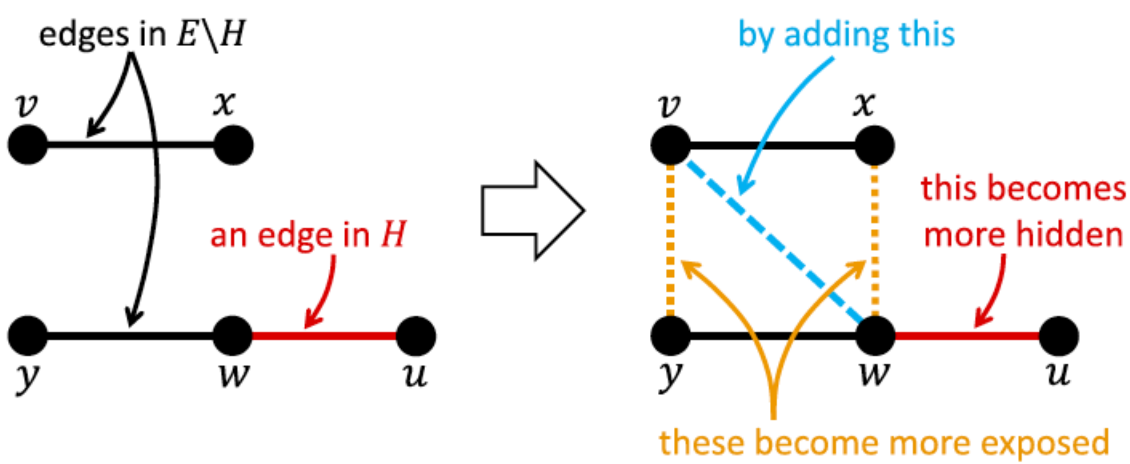


Figure 2. An illustration of the main idea behind the OTC heuristic. Here, the addition of (v, w) creates two open triads: one contains the nodes x, v, w; the other contains v, w, y. In such situations, the similarity scores of (x, w) and (y, v) increase while that of (w, u) could also decrease; see the analysis in *Materials and Methods*.

Reading

- Ch.04 Networks in Their Surrounding Context [NCM]
- Ch.09 Link Prediction [SNA]
- How to hide one's relationships from link prediction algorithms. Waniek, M., et al. Scientific reports'19.