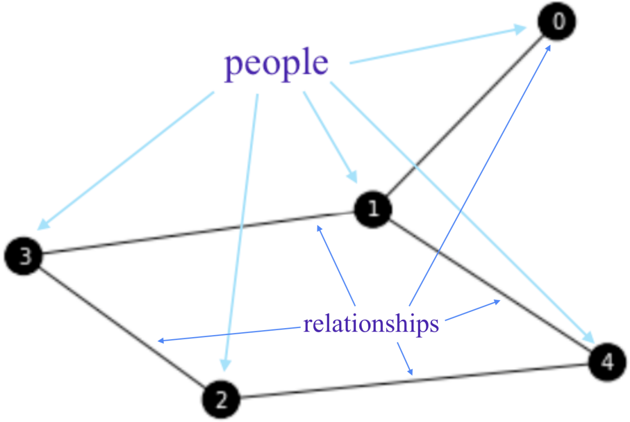
# Link Prediction and Influencer Identification on Weighted Graphs

Raina Wu
Mentor: Professor Laura Schaposnik, University of Illinois at Chicago

MIT PRIMES Conference

October 2023

# Graphs and Social Networks



# Weights as Transmission Probabilities

- Edge weights have the following properties:
  - can represent
    - friendship strength
    - physical proximity
    - frequency of interaction
    - probability of transmission
  - $w(u,v) \in [0,1] \ \forall uv \in E$
  - $w(u,v)$ is approximated by $\frac{\text{\# of interactions}}{\text{units of time}}$
- $T_{u,v}$ is the units of time until $u$ and $v$ interact

**Definition**

Expected transmission time is $d(u,v) = \mathbb{E}[T_{u,v}] = \frac{1}{w(u,v)}$

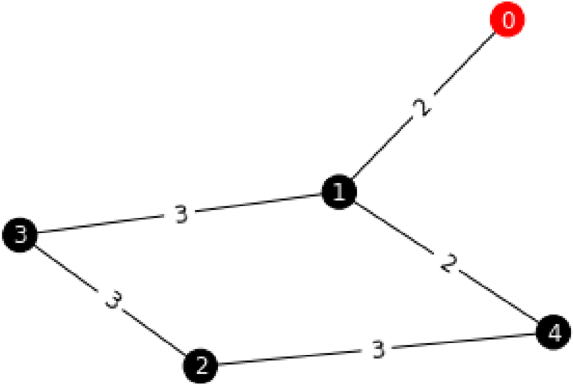
# Modeling Information Diffusion

- How can we model information spread?
  - epidemiological models: an "infection" of information
- Does a disease model (think common cold) always fit?
  - No; peer pressure and social reinforcement exist
- Two general categories: simple contagion (disease) and complex contagion (behavior)

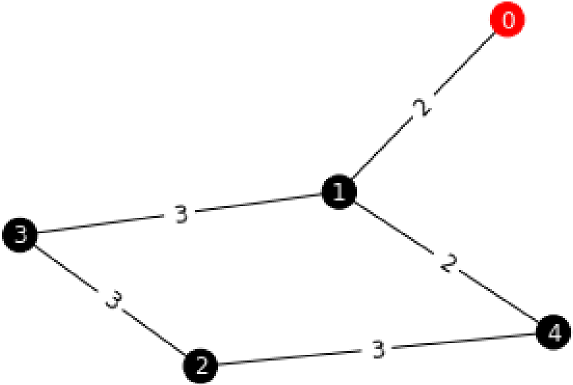
# Simple Contagion

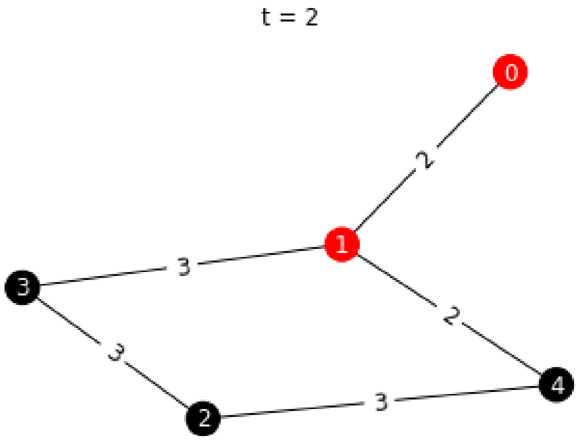
- A single successful interaction is enough to create adoption
  - easily accepted, e.g. conversational topics, facts, the flu
- Each edge $uv$ has a fixed probability $p_{uv}$ of transmission – note that this is just $w(u, v)$
- Again, the expected transmission time between $u$ and $v$ is $d(u, v) = \frac{1}{w(u,v)}$

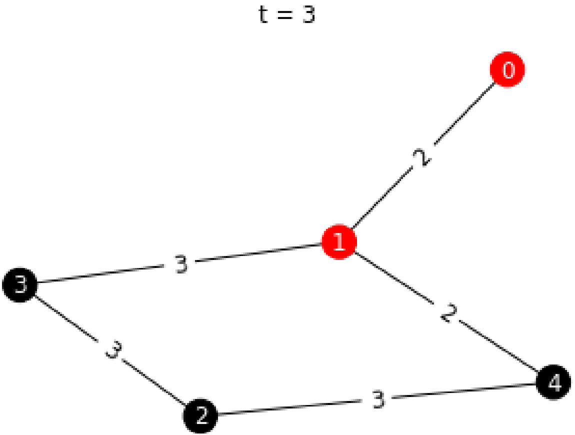
t = 0

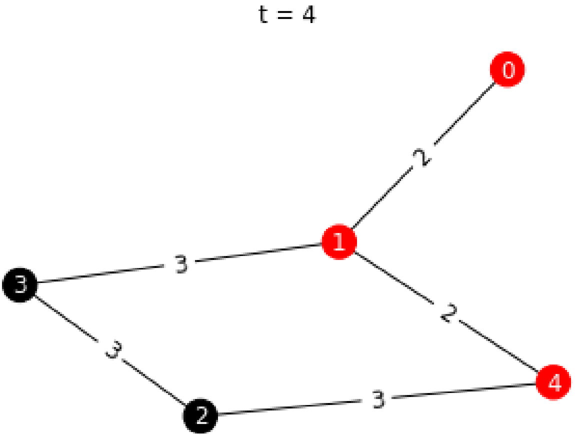


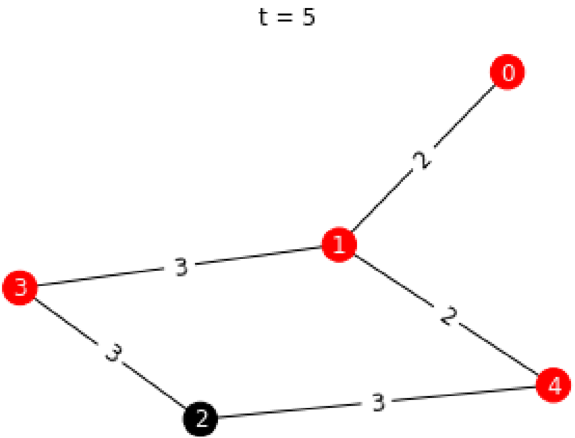
t = 1




t = 2
0
1
2
3
4
2
2
3
3
3


t = 3
0
1
2
3
4
2
2
3
3
3


t = 4
0
1
2
3
4
2
2
3
3
3


t = 5
0
2
1
3
3
2
3
3
2
3
4

# Model Definition

## Definition

Given a set of initially infected nodes $I_0$ in the graph $G = (V, E)$, at time $t$ the set of infected nodes $I_t$ will be

$$I_t = \{v | v \in V \; \exists u : u \in I_0, d_G(u, v) \leq t\}$$

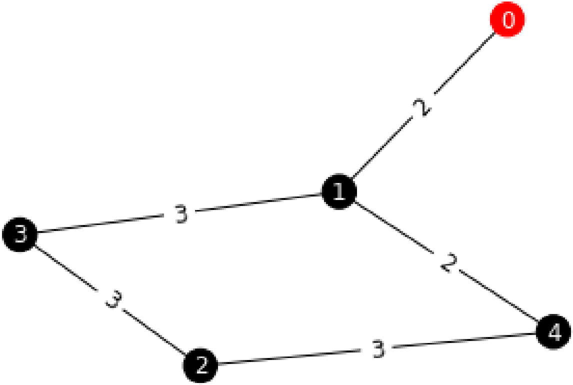
# Complex Contagion

- Multiple successful interactions (reinforcement) needed
  - more difficult topics, e.g. controversial topics, politics, health behaviors
- Often modeled with threshold models

## Definition

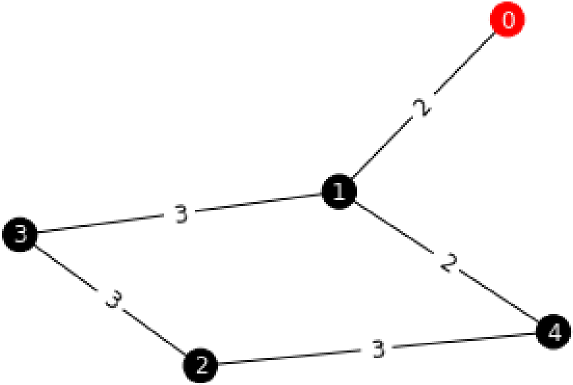
Given an infection value $\theta \in [0,1]$ and infected node set $I_{t-1}$ at time $t-1$, uninfected node $v$ will become infected for time $t$ if $\frac{\sum_{i \in I_{t-1} \cap N(v)} w(i,v)}{\sum_{i \in N(v)} w(i,v)} \geq \theta$. Call $\theta \cdot \sum_{i \in N(v)} w(i,v)$ as its *threshold*.

t = 0



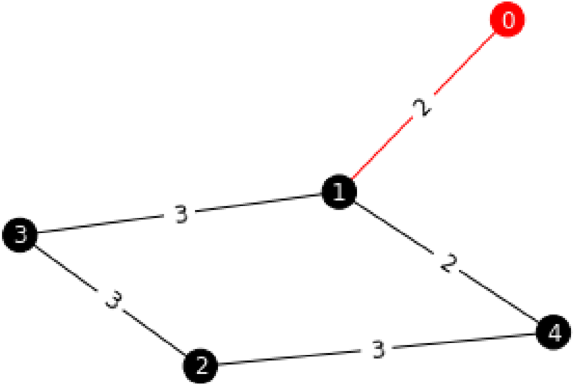

Successful interactions will be shown by red edges

t = 1



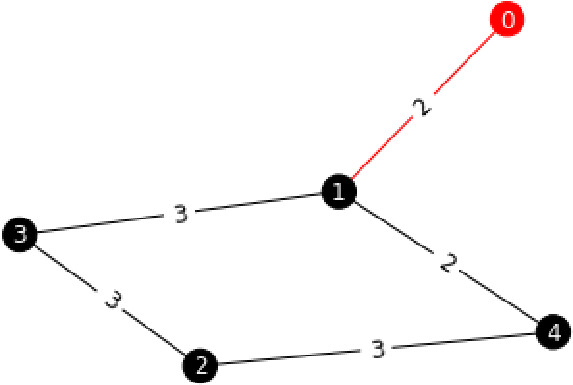

Time elapsed has not surpassed the distance, so edge 01 is not yet red

t = 2




- Even though a successful interaction occurs, no new nodes become infected
- The uninfected endpoint of the red edge at $t = 2$ has threshold $\frac{1}{2} \cdot (\frac{1}{2} + \frac{1}{3} + \frac{1}{2}) = \frac{2}{3}$, and the infected edge only has a weight of $\frac{1}{2}$

t = 7



# Future Networks

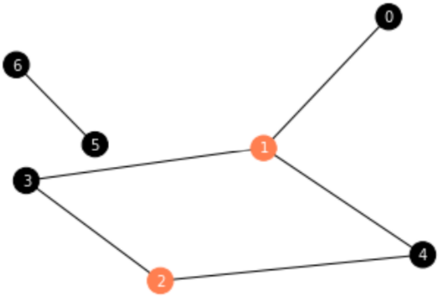
- How can we predict the future of graphs?
- Focus on future edges
- For each pair $u, v \in V, uv \notin E$, we can calculate a similarity score $s_{u,v}$ to estimate probabilities of future connection

# The Common Neighbors Intuition

## Definition

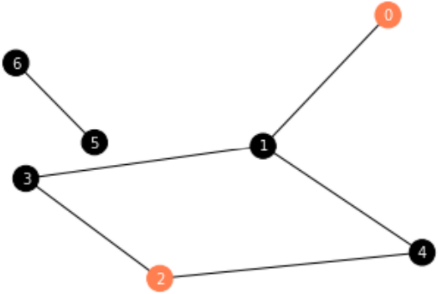
The common neighbors similarity is $s_{u,v}^{CN} = |N(u) \cap N(v)|$.

- Considers first-order neighbors



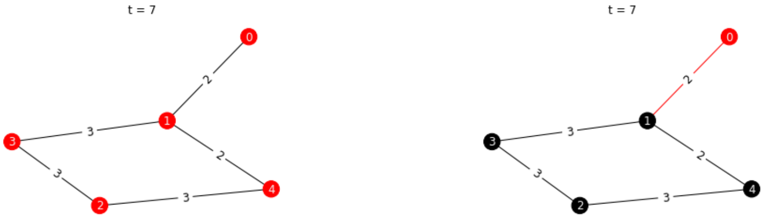
# Variants and Extensions

- Weighted variants consider sums of path length
- Can be extended to second-order neighbors (quasi-local extension)



# Influence Maximization Problem

- Want to choose the $k$ nodes such that influence is maximized
- Influence differs depending on the contagion model:




(a) Simple Contagion with initial infected node 0

(b) Complex Contagion with initial infected node 0

# Heuristic Centrality Metrics

- Primarily concerned with searching for a single influencer ($k = 1$)
- General categories:
  - local measures, e.g. degree centrality
  - iterative measures, e.g. PageRank, LeaderRank, coreness
  - global measures, e.g. eigenvector centrality
- For a centrality metric, the top-scoring node is its "influencer"

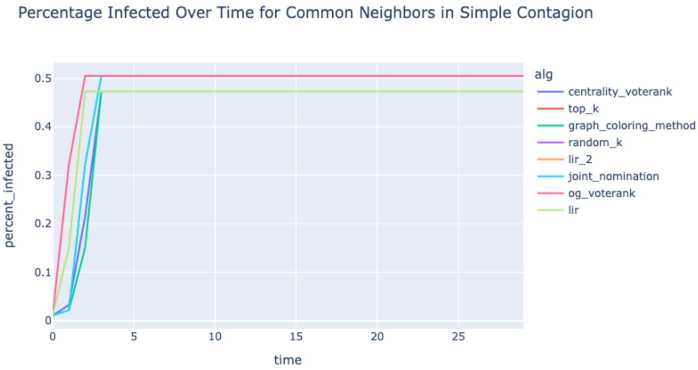
# Choosing $k$ Nodes

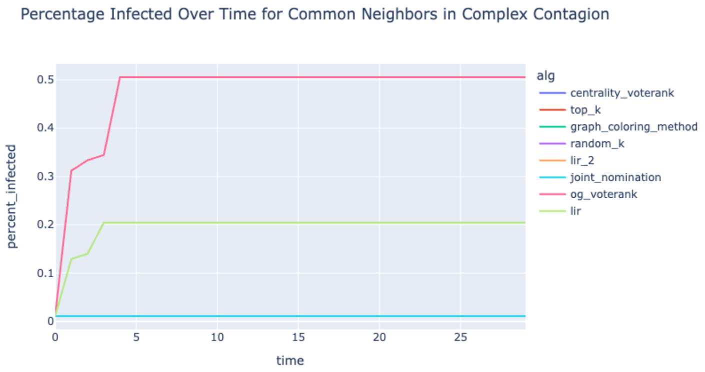
- Chooses a team instead of an individual
- Some use recursion around neighborhoods
  - e.g. VoteRank, where nodes vote for neighbors
- Can also incorporate centrality metrics after reducing redundancy
  - e.g. graph coloring, which separates the graph into independent sets before running centrality

## Predicted Graphs

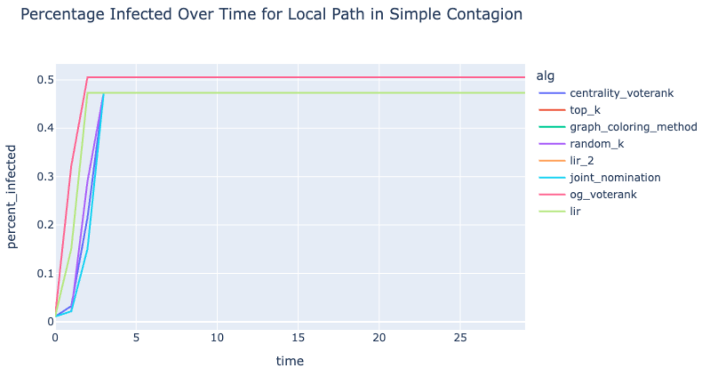
- Steps to predict future influencers/groups of $k$ nodes:
  1. Given a graph, randomly take 90% of its edges as a starting graph
  2. Do link prediction on the starting graph and calculate similarity scores for each pair of nodes $(u, v)$
  3. If $s_{u,v} \neq 0$, normalize it into a probability of existence, which becomes a probability of transmission
  4. Run centrality and top $k$ algorithms on the predicted graph to find a set of predicted $k$ nodes
  5. Test the set found on the original graph to measure final number of nodes infected

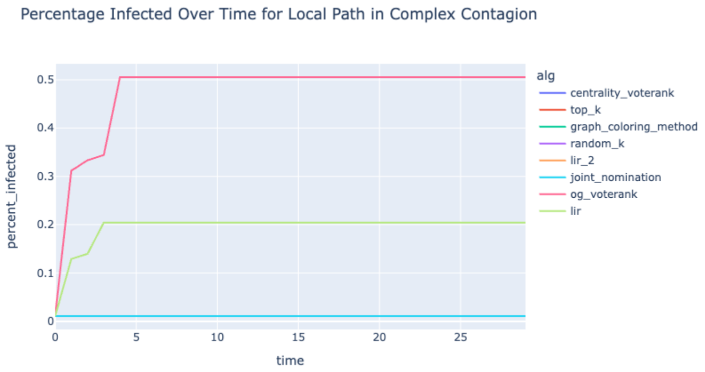
# Examples When Run on Graphs






Pink: VoteRank; Green: LIR, LIR-2, Blue: rest


Percentage Infected Over Time for Local Path in Simple Contagion
percent_infected
0
0.1
0.2
0.3
0.4
0.5
0
5
10
15
20
25
time
alg
centrality_voterank
top_k
graph_coloring_method
random_k
lir_2
joint_nomination
og_voterank
lir




Pink: VoteRank; Green: LIR, LIR-2, Blue: rest

# Applications

- Can be applied to:
  - advertising/marketing
  - social movement analysis
  - epidemiology
  - rumor propagation
  - media propaganda
- Help with prevention and planning

# Acknowledgements

- I would like to thank my mentor, Prof. Laura Schaposnik, for her guidance and encouragement throughout the project.
- I am grateful to the MIT PRIMES-USA Program, Dr. Tanya Khovanova, Dr. Slava Gerovitch, and Prof. Pavel Etingof for making such a wonderful research opportunity.
- My parents

# References


- Amit Goyal, Francesco Bonchi, and Laks Lakshmanan. Learning influence probabilities in social networks. volume 241-250, pages 241–250, 02 2010.
- Byungjoon Min and Maxi Miguel. Competing contagion processes: Complex contagion triggered by simple contagion. Scientific Reports, 8, 07 2018.
- Damon Centola and Michael Macy. Complex contagions and the weakness of long ties. American Journal of Sociology, 113(3):702–734, 2007.
- Paulo Shakarian, Abhinav Bhatnagar, Ashkan Aleali, Elham Shaabani, and Ruocheng Guo. The Independent Cascade and Linear Threshold Models, pages 35–48. 01 2015.
- Tsuyoshi Murata and Sakiko Moriyasu. Link prediction of social networks based on weighted proximity measures. volume 85-88, pages 85–88, 12 2007.
- Tao Zhou, Linyuan Lü, and Yi-Cheng Zhang. Predicting missing links via local information. The European Physical Journal B, 71(4):623–630, oct 2009.

- Meng Bai, Ke Hu, and Yi Tang. Link prediction based on a semi-local similarity index. Chinese Physics B, 20(12):128902, dec 2011.
- Furqan Aziz, Haji Gul, M. Irfan Uddin, and Georgios Gkoutos. Path-based extensions of local link prediction methods for complex networks. Scientific Reports, 10, 11 2020.
- V. Batagelj and M. Zaversnik. An o(m) algorithm for cores decomposition of networks, 2003.
- P. D. Karampourniotis, B. K. Szymanski, and G. Korniss. Influence maximization for fixed heterogeneous thresholds. Scientific Reports, 9(1), apr 2019.
- Jian-Xiong Zhang, Duan-Bing Chen, Qiang Dong, and Zhi-Dan Zhao. Identifying a set of influential spreaders in complex networks, 2016.
- J. Leskovec, J. Kleinberg and C. Faloutsos. Graph Evolution: Densification and Shrinking Diameters. ACM Transactions on Knowledge Discovery from Data (ACM TKDD), 1(1), 2007.