Influence Maximization in Online Social Networks

Resa Brockman

University of Minnesota, Morris

April 30, 2016

The What and Why of Influence Maximization (IM)

- Finding the `x` number of the most influential people (*seed nodes*) in a network
- Why? **Marketing**

The Goal

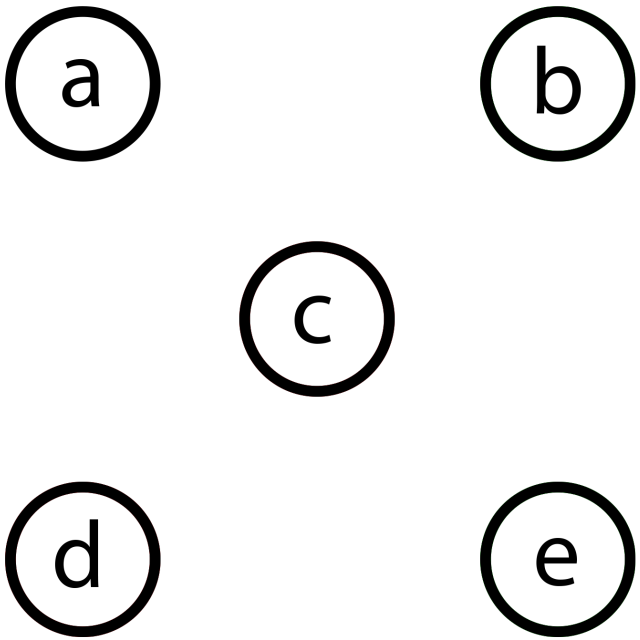
To spread information (*influence*) to as large a portion of a network as possible.

Outline

The Setup

The Setup: Influence Graphs

- A network can be represented as an *influence graph*

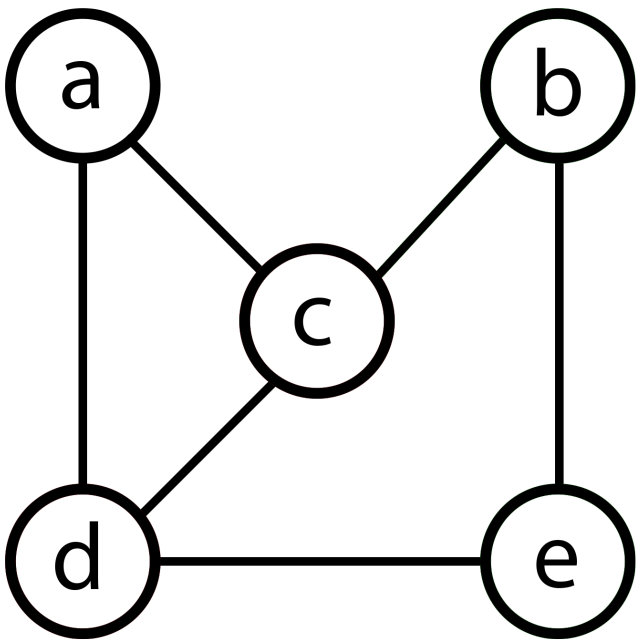


- *Nodes* are used to represent people

Graphic courtesy of Michelle King

The Setup: Influence Graphs

- A network can be represented as an *influence graph*

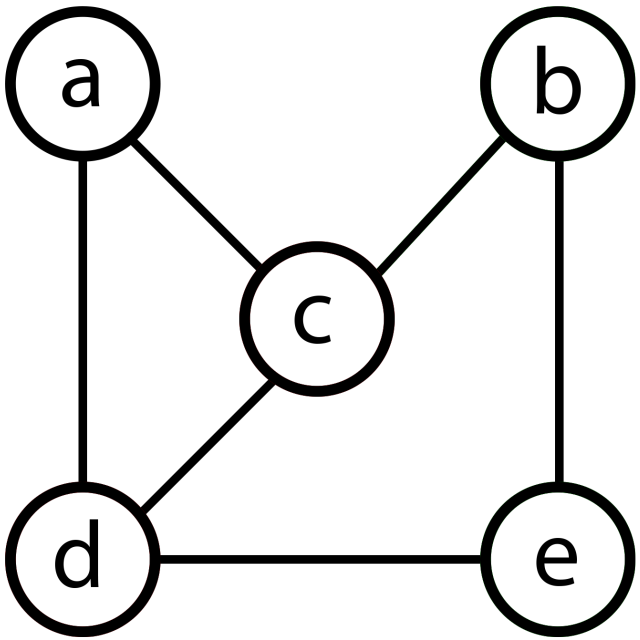


- *Nodes* are used to represent people.
- Nodes are connected by *influence probabilities*

Graphic courtesy of Michelle King

The Setup: Influence Graphs

- A network can be represented as an *influence graph*

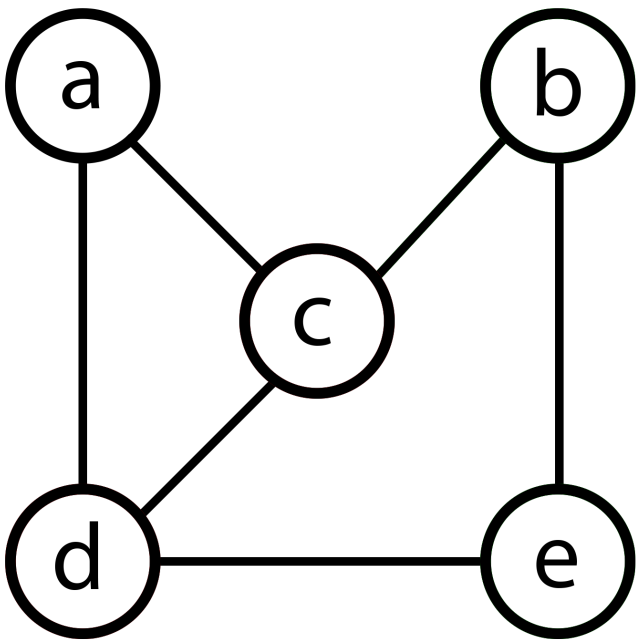


- *Nodes* are used to represent people.
- Nodes are connected by *influence probabilities*
- They can be **asymmetric**

Graphic courtesy of Michelle King

The Setup: Influence Graphs

- A network can be represented as an *influence graph*

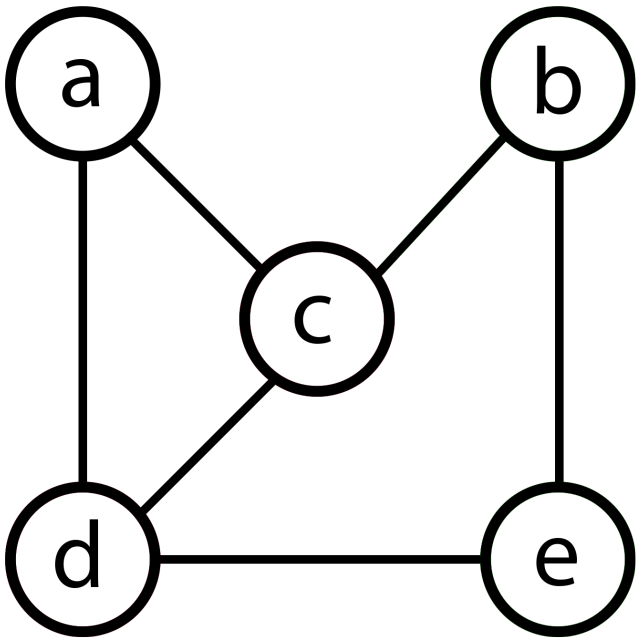


- *Nodes* are used to represent people.
- Nodes are connected by *influence probabilities*
- They can be **asymmetric**
- Or **symmetric**

Graphic courtesy of Michelle King

Independent Cascade Model

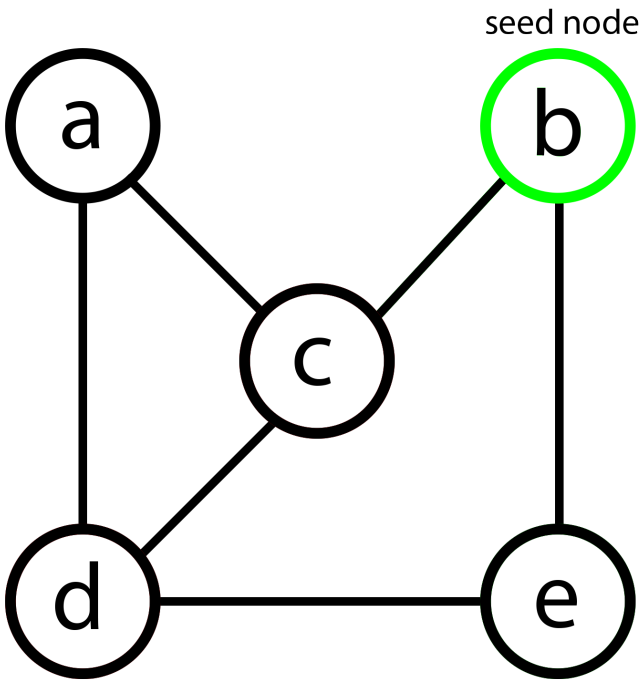
- Also called an *Event Cascade Model* or *Diffusion Model*



Graphic courtesy of Michelle King

Independent Cascade Model

- Also called an *Event Cascade Model* or *Diffusion Model*

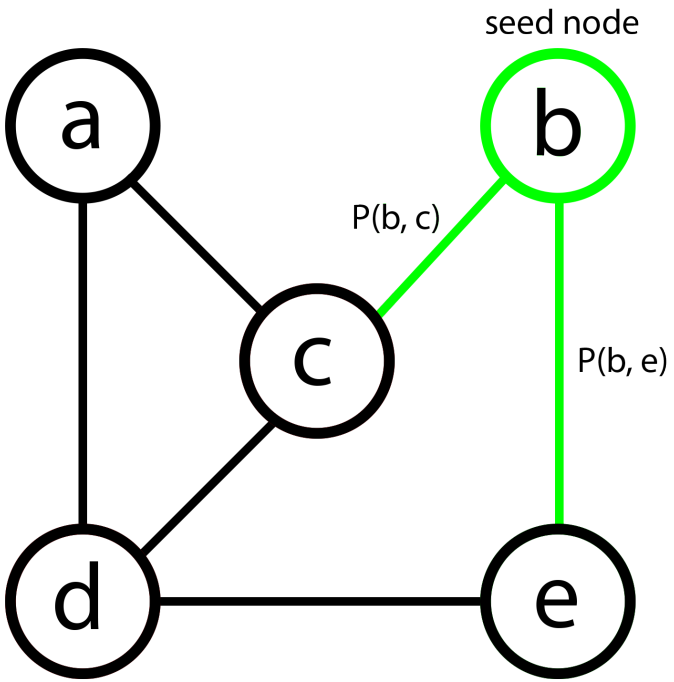


1. Some seed nodes are *activated*

Graphic courtesy of Michelle King

Independent Cascade Model

- Also called an *Event Cascade Model* or *Diffusion Model*

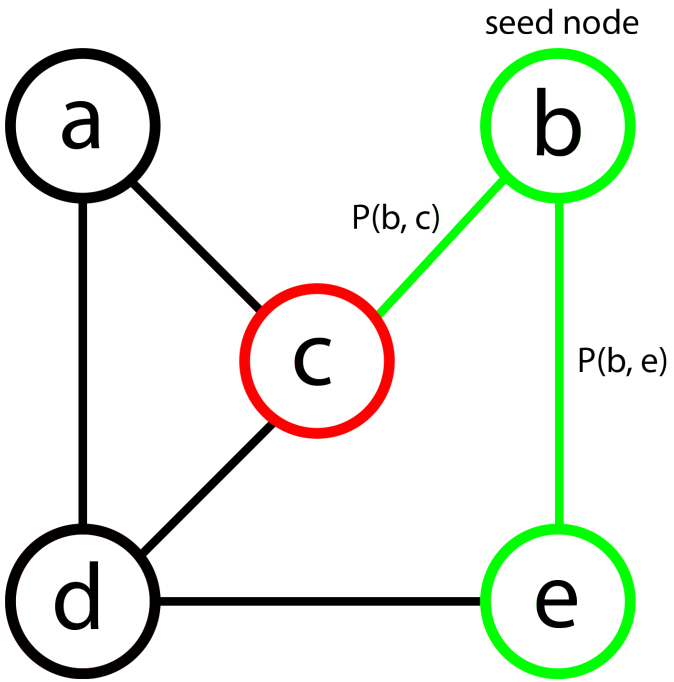


1. Some seed nodes are *activated*
2. Each activated node tries to activate connected nodes with the connected influence probability

Graphic courtesy of Michelle King

Independent Cascade Model

- Also called an *Event Cascade Model* or *Diffusion Model*

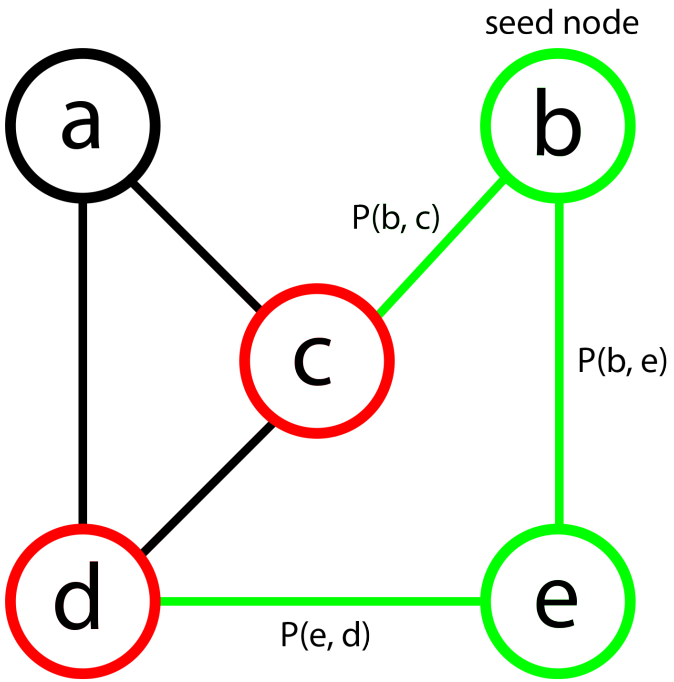


1. Some seed nodes are *activated*
2. Each activated node tries to activate connected nodes with the connected influence probability

Graphic courtesy of Michelle King

Independent Cascade Model

- Also called an *Event Cascade Model* or *Diffusion Model*



1. Some seed nodes are *activated*
2. Each activated node tries to activate connected nodes with the connected influence probability
3. Final *influence spread* is determined

Graphic courtesy of Michelle King

Early Solutions

- Solutions were adopted and modified from information diffusion research in the social sciences
- As such, IM has been formulated as a combinatorial optimization problem since 2003
- Early solutions were simple and effective, but extremely computationally expensive

Combinatorial Optimization

Finding an optimal object from a finite set of objects. In many such problems, exhaustive search is not feasible.

Two-phase Influence Maximization (TIM)

- In use since 2014
- Returns a solution equivalently good to the best (in terms of accuracy) previous algorithm (2003)
- Near linear expected time under the independent cascade model
- TIM requires less than one hour to process a network with 41.6 million nodes and 1.4 billion edges

TIM Summary

1. TIM is given:
   - The social network (G)
   - The desired number of seed nodes (k)
2. Algorithm 1 determines the expected spread of influence per node (t)
3. Using t, TIM calculates how many nodes it needs to sample for the most optimal solution (θ)
4. Algorithm 2 randomly samples θ nodes and chooses the best k seed nodes out of those

Reverse-Reachable (RR) Sets

- The set of nodes that can reach a given node
- Found by removing edges with 1-probability of activation between the two nodes
- If an edge is successfully removed, it is added to the RR set

Algorithm 1: Determine the Expected Spread of Influence

- Sample $log_2 n - 1$ nodes
- For each sampled node, determine its RR set (spread)
- Sum the spreads of each node, and divide by number of nodes sampled (average spread)
- Return the average spread (t)

Calculate θ

- Recall that θ is the number of nodes that should be sampled for a good but computationally reasonable result
- Calculating θ is one of the most important contributions of TIM to improving IM accuracy
- Generalizing some very complex math:
  $\theta \geq \frac{n}{m} t \ / \ \textit{maximum expected spread of a } k - \textit{sized node set}$
- In actuality, using an estimation **smaller** than $\frac{n}{m} t$ in the numerator provides equally good and computationally less expensive results

Algorithm 2: Return k Seed Nodes

- This is the second algorithm (second phase) of TIM
- Creates a set of θ RR-sets (R)
- Choose the node from R set with the largest spread
- Remove nodes from R that cover the same nodes
- Continue until k nodes have been chosen

Incomplete Influence Data

- Influence probabilities come from users' logs of past activities
- It is not uncommon for influence probabilities to be missing or unavailable
- Usually, a given influence probability used for all missing data
- Leads to poorly chosen seed nodes

Multiple -Trial Solution

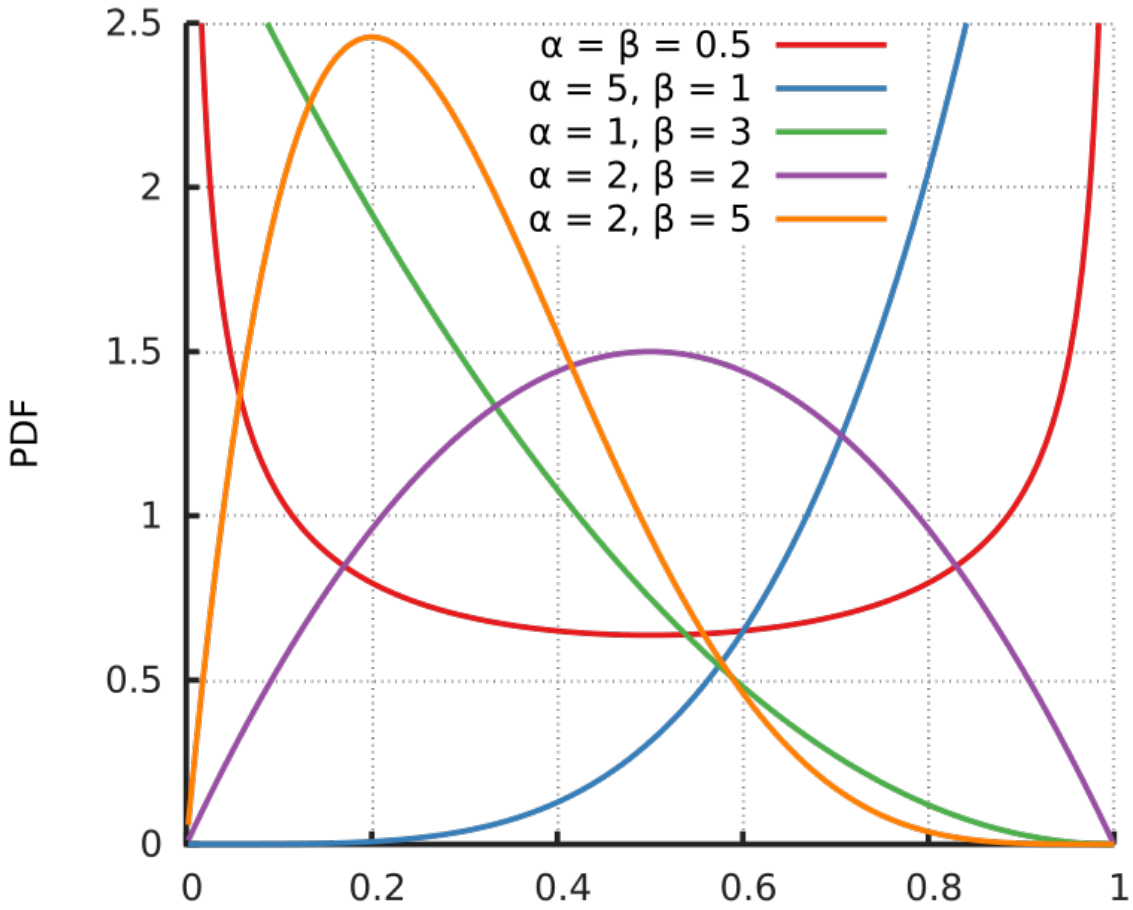
- New as of 2015
- Researchers named this approach *Online Influence Maximization* (OIM)
- Can be used with any existing IM algorithm (TIM used)
- Requires real-world trials (budget permitting)

OIM Summary

1. Assign influence probabilities where missing
2. Choose seed nodes using existing IM algorithm
3. Run real-world trial, collect user feedback data
4. Update influence probabilities
5. Repeat according to time-frame or budget

Influence Probabilities - Using Beta Distribution

- Probability can range from 0 to 1
- Beta Distribution has two parameters: $B(\alpha, \beta)$
- α will be the success parameter, β the failure parameter



OIM Setup

- For missing influence probabilities α and β are both initially set to 1: $B(1, 1)$ (uniform distribution)
- For existing influence probabilities, set α and β accordingly
- Use IM algorithm (e.g. TIM) to find k seed nodes
- Run real-world trial

Update Influence Probabilities

- Feedback information from the trial consists of:
  - The set of ultimately activated nodes
  - The set of edge activation attempts and outcomes (successful/unsuccessful)
- This is used to update the influence probabilities
  - If an activation attempt was a **success**, add 1 to α: $B(\alpha + 1, \beta)$
  - If an activation attempt was a **failure**, add 1 to β: $B(\alpha, \beta + 1)$

Repeat

- Using the updated probabilities, new seed nodes can be chosen, and another real-world trial can be run.
- This can continue as long as:
  - The budget for trials does not run out
  - The improvements made each trial are not trivial
  - The marketing campaign continues.
- OIM is best used in networks where feedback information and activation success is easy to determine
- Micro-blogging networks are ideal for this (and most IM work)

Everybody Wants to be a Trendsetter

- Another way to find and define influential people in a network: *trendsetters*
- A trendsetter is defined by two things:
  1. Having a specific area of interest or expertise
  2. Adopting new ideas or trends in this area before most others (Specifically trends that eventually become very popular)
- Trendsetters can only be found after some trend of interest has become popular
- Trendsetters are found with a ranking algorithm

PageRank

- PageRank was developed by Larry Page, one of the founders of Google
- PageRank counts the number and quality of links to a page to estimate its importance
- First algorithm used to order Google search results
- PageRank can be generalized and used on any graph or network

Generalized Pagerank Algorithms

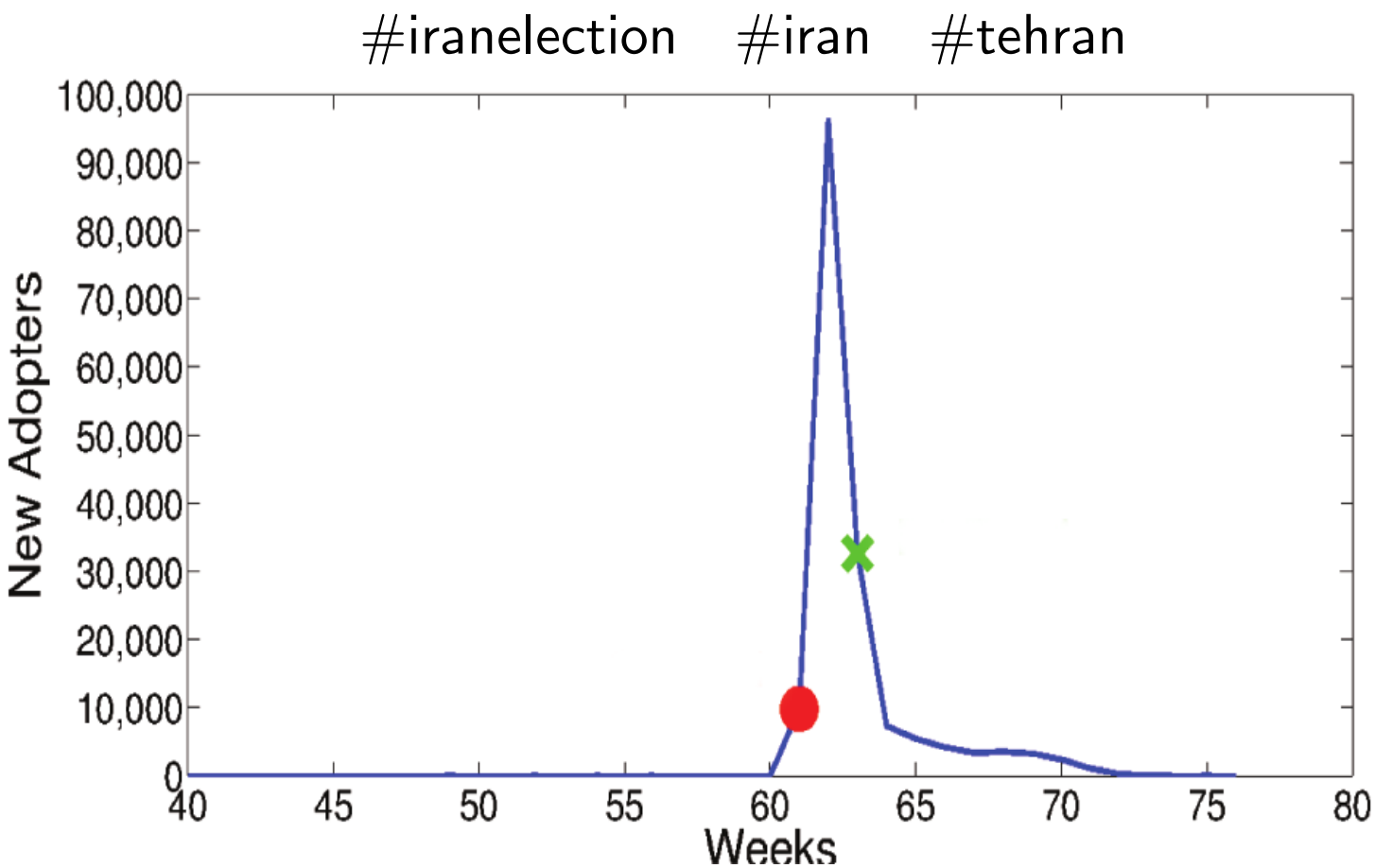
- Twitter uses a pagerank algorithm to recommend accounts to follow
- A pagerank algorithm has been used to rank streets in order of popularity (high traffic)
- Pagerank algorithms can be used to determine the most essential species in an ecosystem
- To find trendsetters, a pagerank algorithm is combined with time information

Ranking Trendsetters

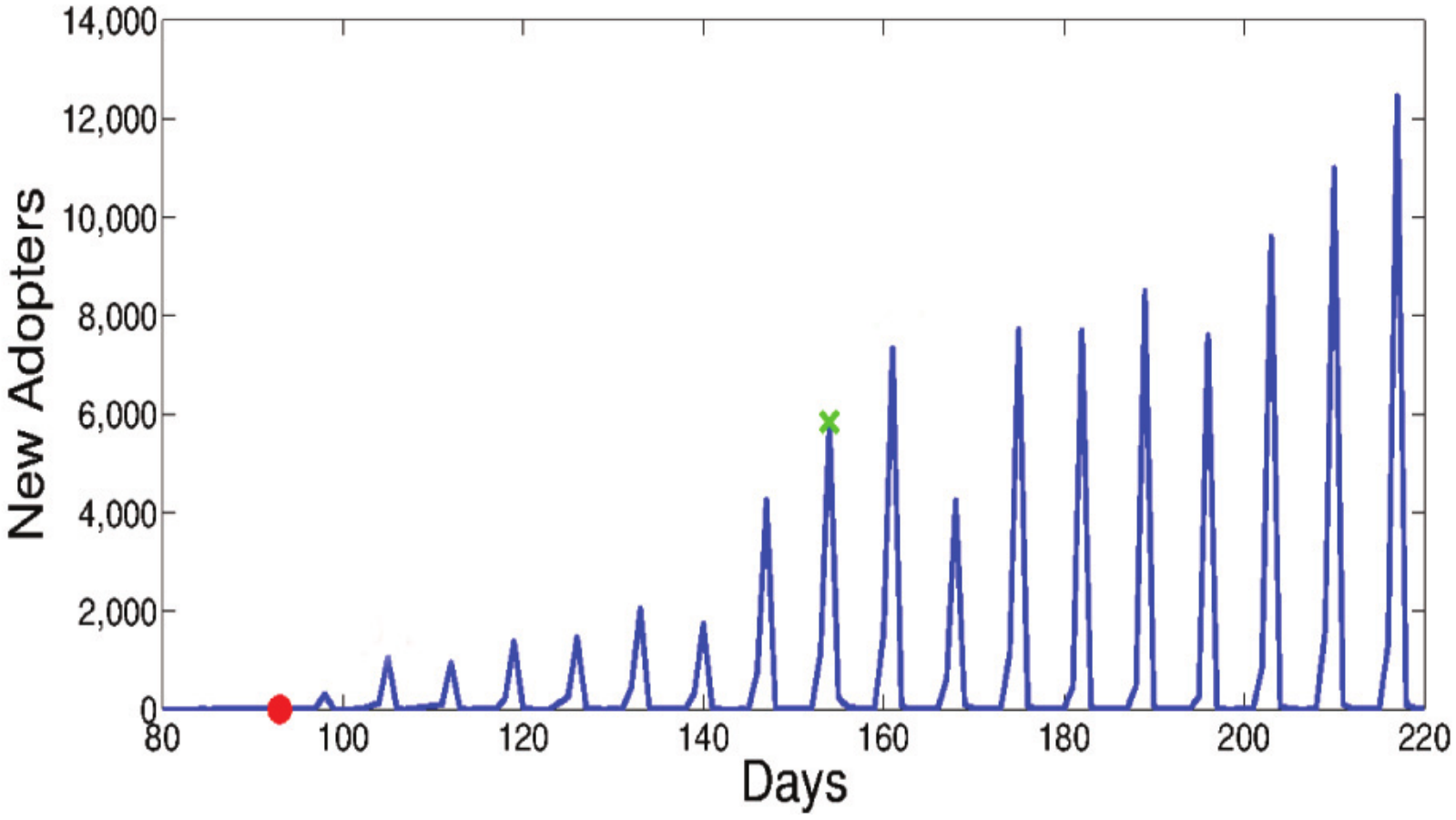
- Approach developed using Twitter

1. Define a trend using hashtags
2. For each user, determine:
   - How many trend hashtags the user used
   - How many followers the user has
   - When the user began using the hashtags (new)
3. Use modified pagerank to rank users
4. Results can be compared to previous Twitter trendsetter ranking results

2009 Iran Election Timeline



#MusicMonday



Summary

Summary

- TIM is the best influence maximization algorithm

Summary

- TIM is the best influence maximization algorithm
- Missing information can still yield good influence maximization results

Summary

- TIM is the best influence maximization algorithm
- Missing information can still yield good influence maximization results
- There is more than one way to determine who is influential in a network

Summary

- TIM is the best influence maximization algorithm
- Missing information can still yield good influence maximization results
- There is more than one way to determine who is influential in a network
- Everything is a marketing tool in the end, even our friends

Acknowledgments

Thank you KK, Elena, Michelle King, computer science faculty, and everyone who came to see the presentations today!

Questions?