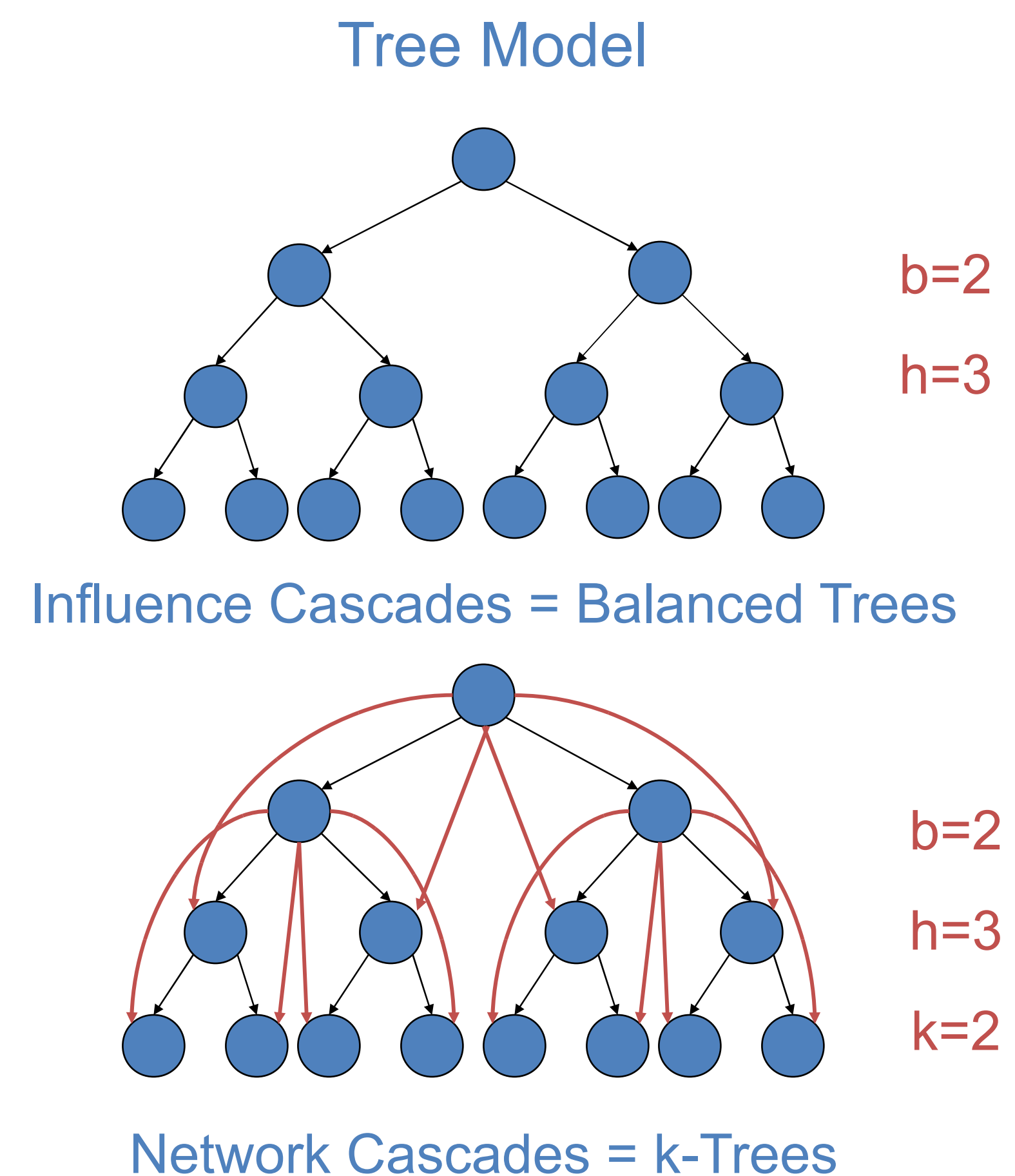
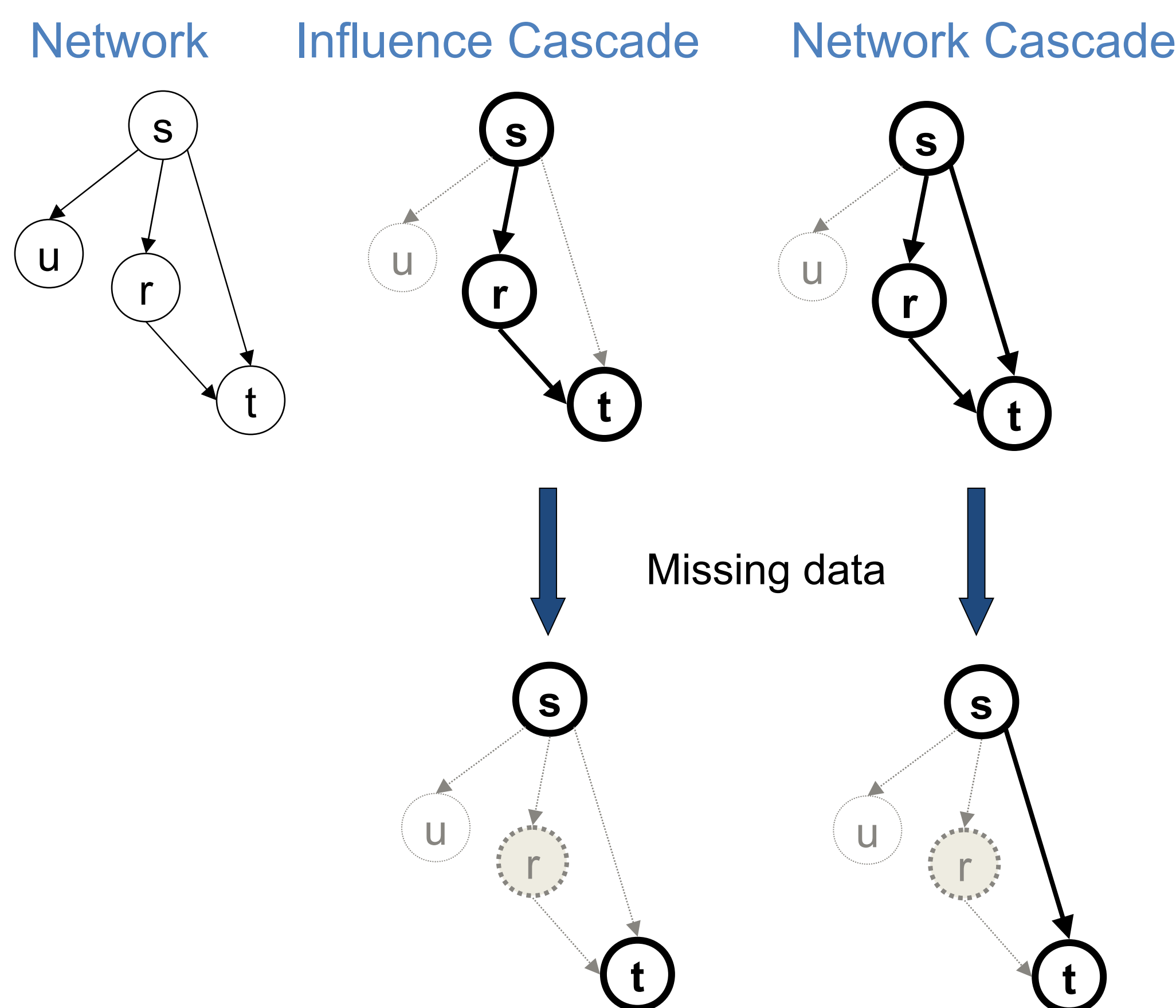


Goals

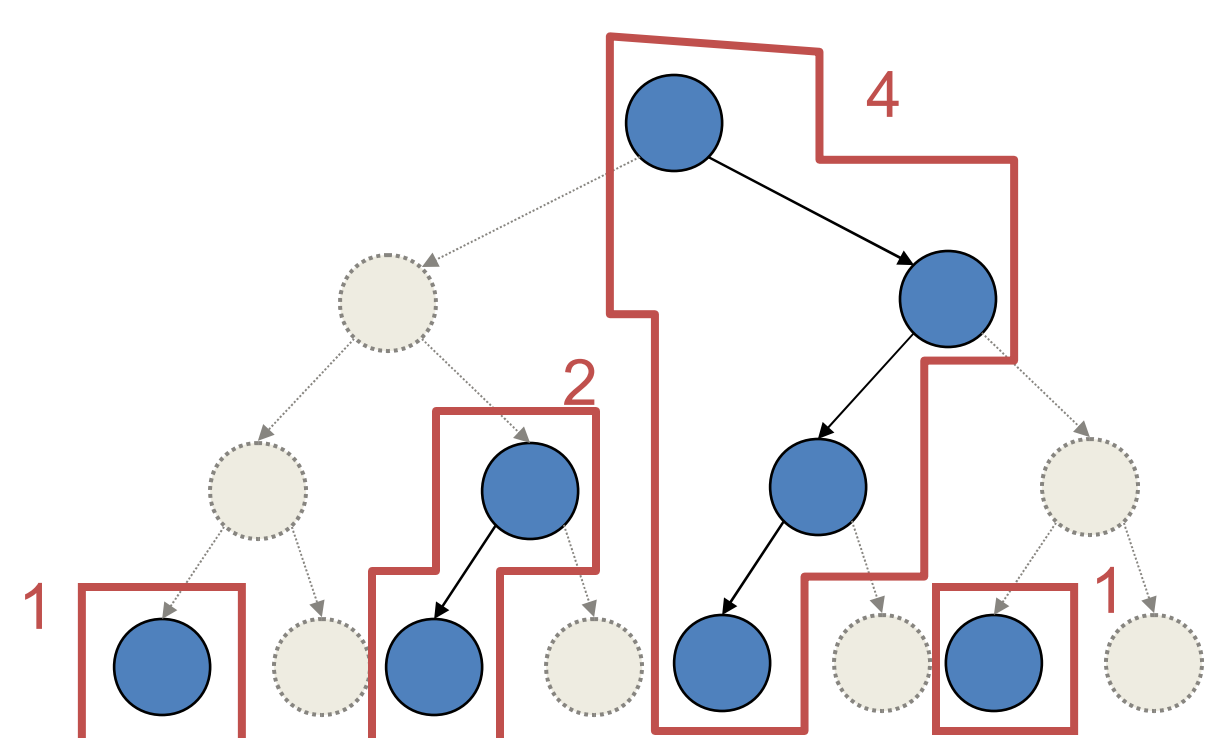
- Study effect of missing data on the inferences about cascading behavior
- Analyze the interplay between the effect of missing data and the structure of the underlying social network

Contributions

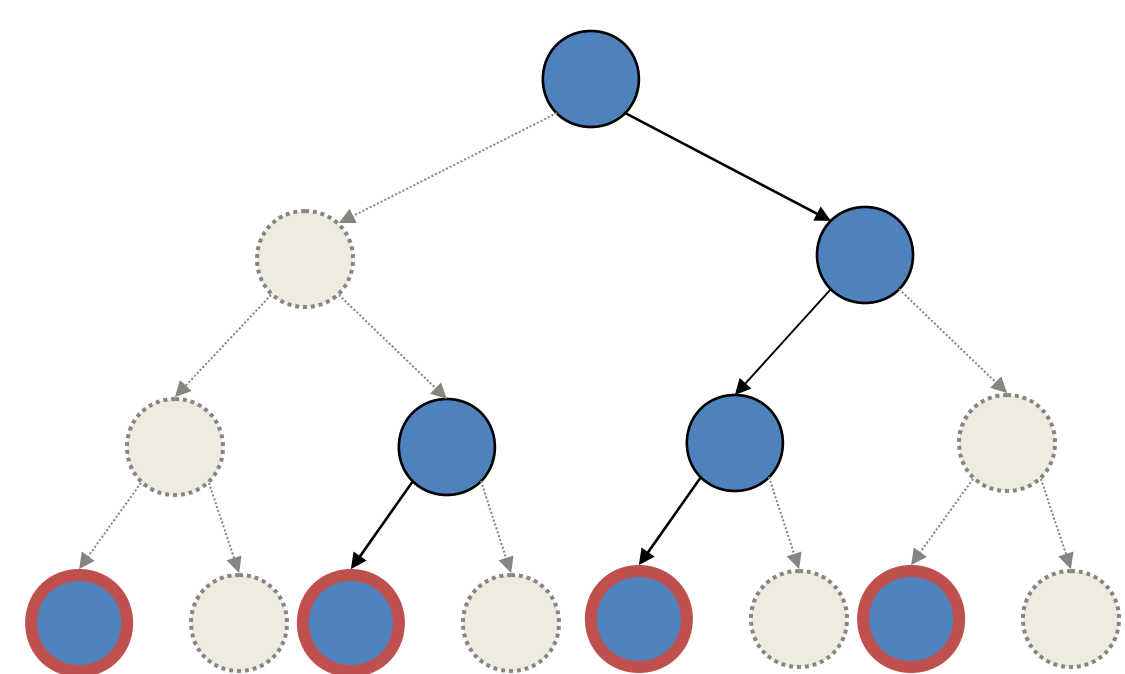
- Formulate a mathematically tractable model of the structure of information cascades
- Show that we can reliably detect and correct for the biases arising from the missing data
- Evaluate our data on retweet cascades over complete Twitter network



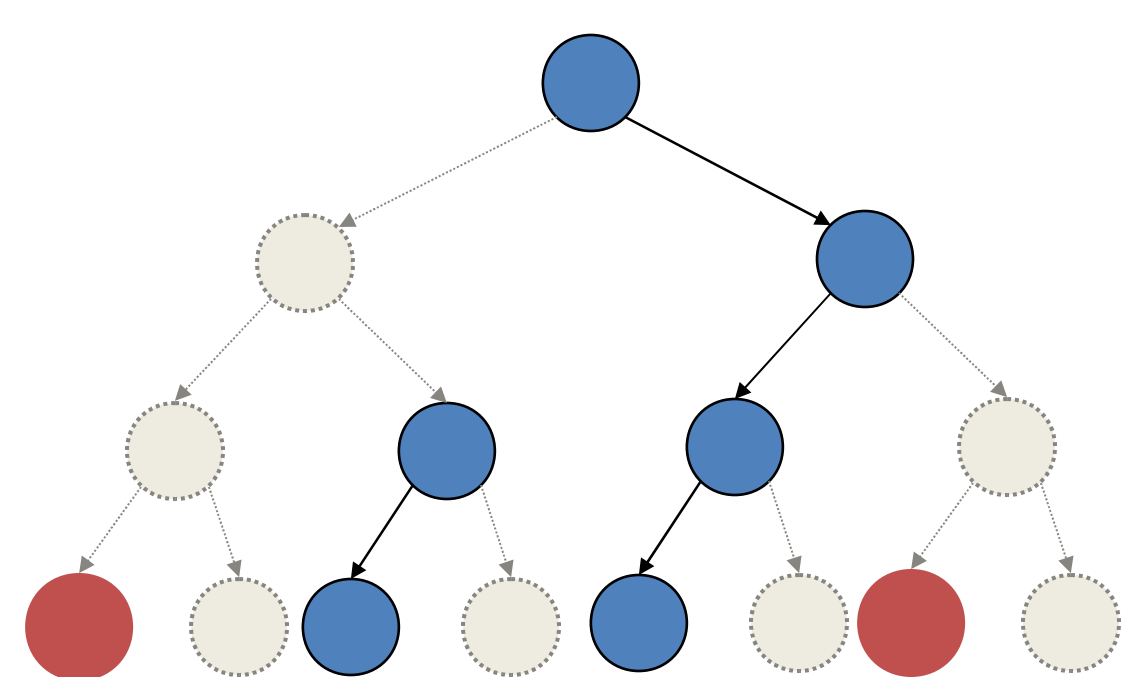
Tractable Properties of Interest



Number and Size of Connected Components



Leaves



Number of Isolated Nodes

Twitter Data Set

Network

- 70 million nodes
- 2 billion edges

Tweets

- Over 1 billion tweets
- Over 4 thousand retweet cascades extracted from mentions of most popular URLs

Parameter Estimation

Are Trees a Good Model for Cascades?

- Calculate average out-degree, in-degree, depth from real data (parameters b , k , h)
- Fit observed data to analytical expressions and compare estimated parameters to observed parameters

How Do We Use Estimated Parameters?

- Can use b , k , h to estimate properties of real cascades
- Can estimate fraction of missing data

Preliminary Results

