Exceptional service in the national interest

Sandia National Laboratories

SAND2016-11644C

A DIFFUSION MODEL FOR MAXIMIZING INFLUENCE SPREAD IN LARGE NETWORKS

Tu-Thach Quach and Jeremy D. Wendt

8th International Conference on Social Informatics (SocInfo2016)

U.S. DEPARTMENT OF ENERGY

NNSA National Nuclear Security Administration

This work was supported by the Laboratory Directed Research and Development program at Sandia National Laboratories, a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000. SAND NO. 2016-XXXX C.

WHY INFORMATION SPREAD?

We need to better understand how information flows online

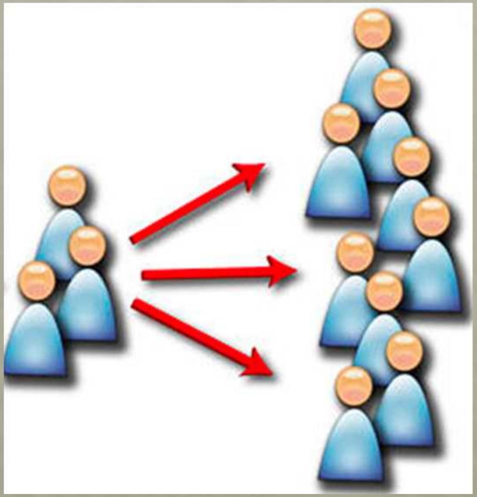
Find the Influencers

- Given a graph, identify the most influential nodes
  - Requires real-world diffusion data and a diffusion model

- Which diffusion model to use?
  1. Should match real-world data
  2. Parameters obtained from real-world data (but good w/o)
  3. Computationally efficient for massive networks

FIND MAXIMIZING SEEDS

- Given a graph, and a diffusion model, find the seed nodes that maximize the diffusion score



- Previous work
    - Independent Cascade and Linear Threshold (Kempe, 2003)
    - Probabilistic Voter indicates highest degree (Even-Dar, 2011)

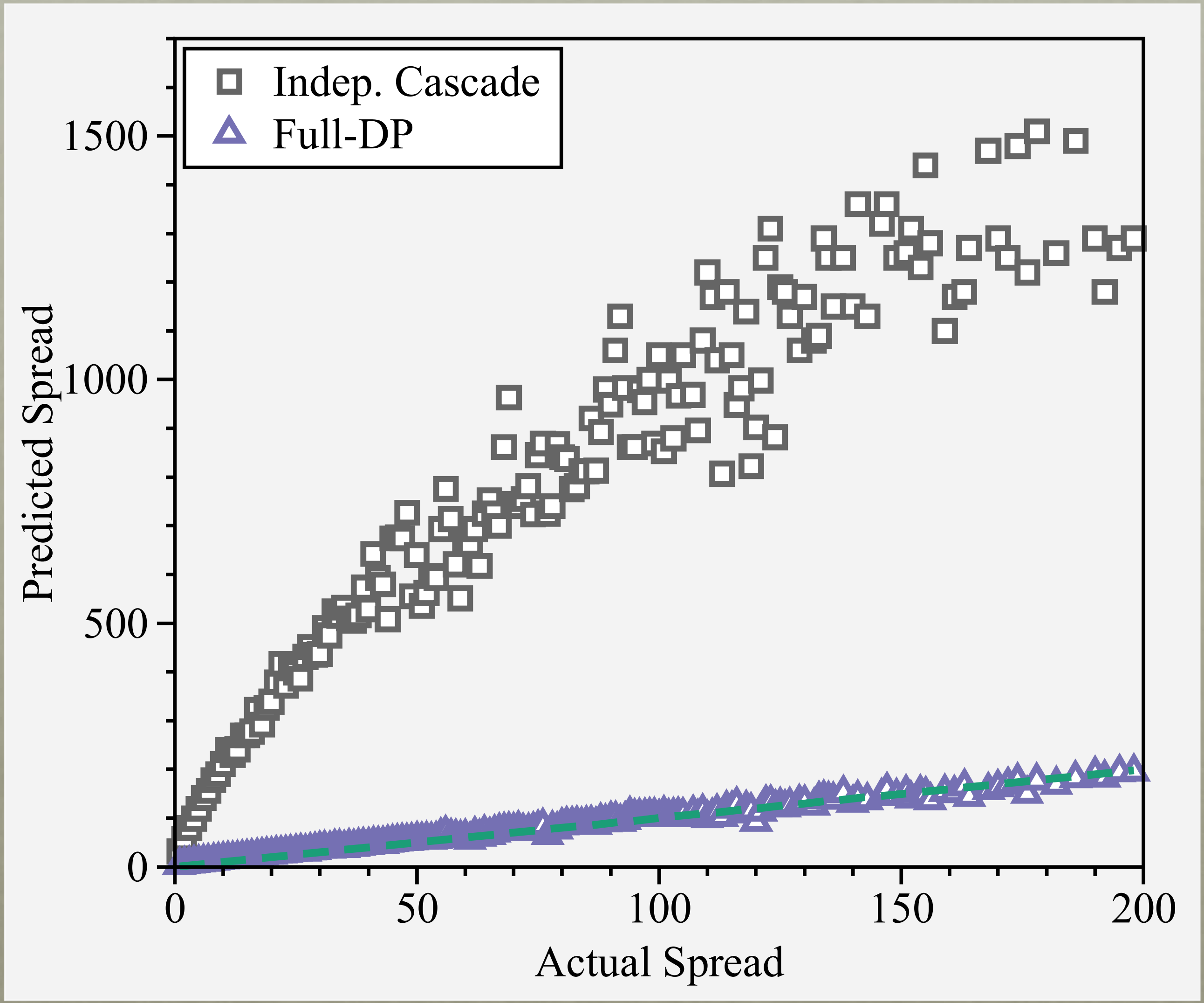
FORWARD PROPAGATION

- A modification of Belief Propagation that preserves directed influence
  - Belief Propagation passes update messages in both directions along edges in a graph
  - In Forward Propagation, messages pass only downstream
- Requires per-node and per-edge functions
  - Per-node may be learned from real-world data
  - Per-edge based on node in-degree
- Results in each node's likelihood of adoption
  - Diffusion score is sum of all nodes' likelihoods
- Implementation details in the paper
  - Available at https://github.com/algorithmfoundry/Foundry
  - Details on BP (Yedidia, 2001)

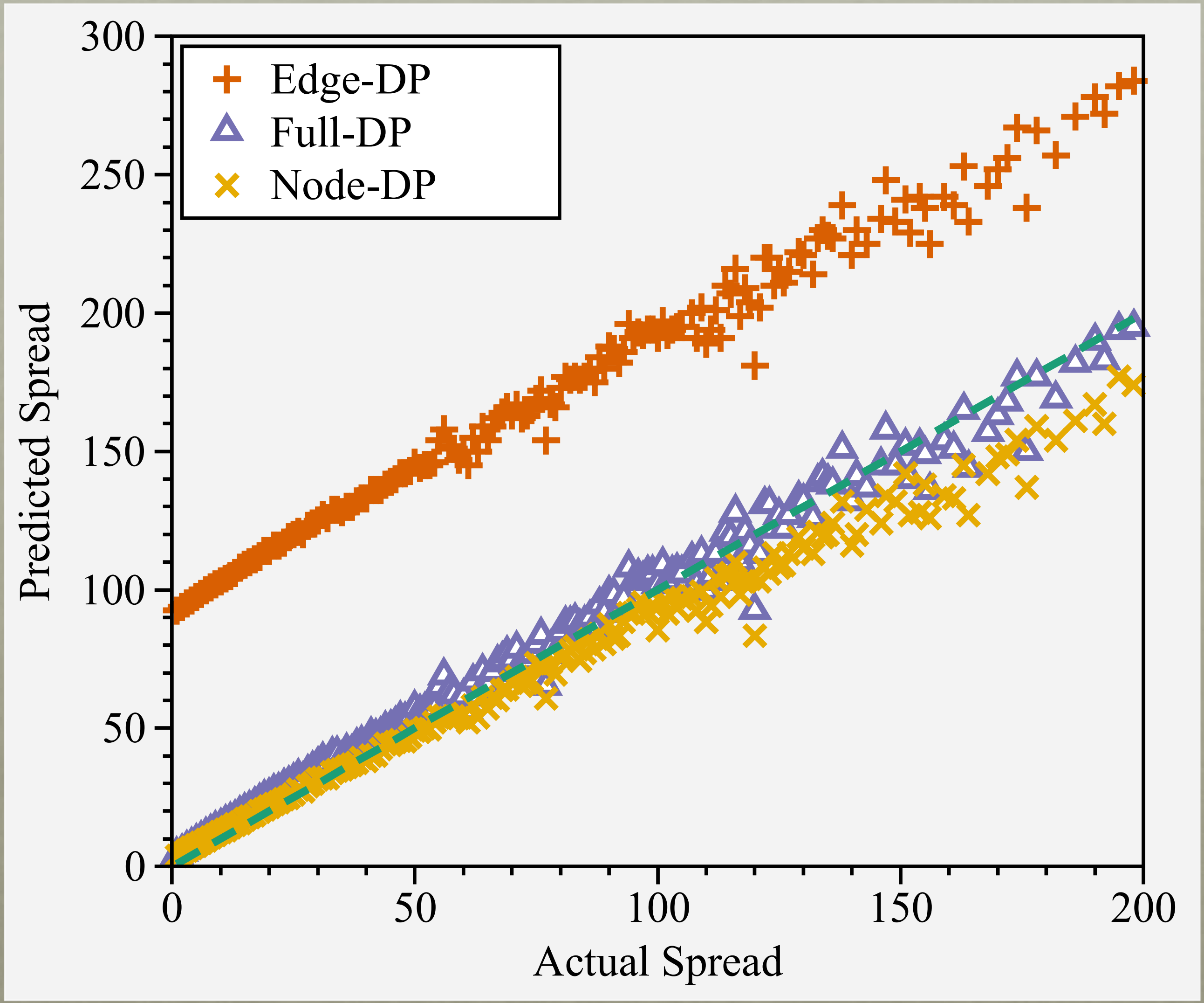
TEST 1: MATCH DATA

- Datasets
    - Flixster movie review propagation
        - 800K nodes; 12M edges
    - Epinions product review propagation
        - 18K nodes; 1.2M edges
- Models
    - Independent Cascade (Kempe, 2003)
    - Directed Propagation
        - degree-weight per-edge; learned per-node (Full-DP)
        - degree-weight per-edge; constant per-node (Edge-DP)
        - constant-weight per-edge; learned per-node (Node-DP)

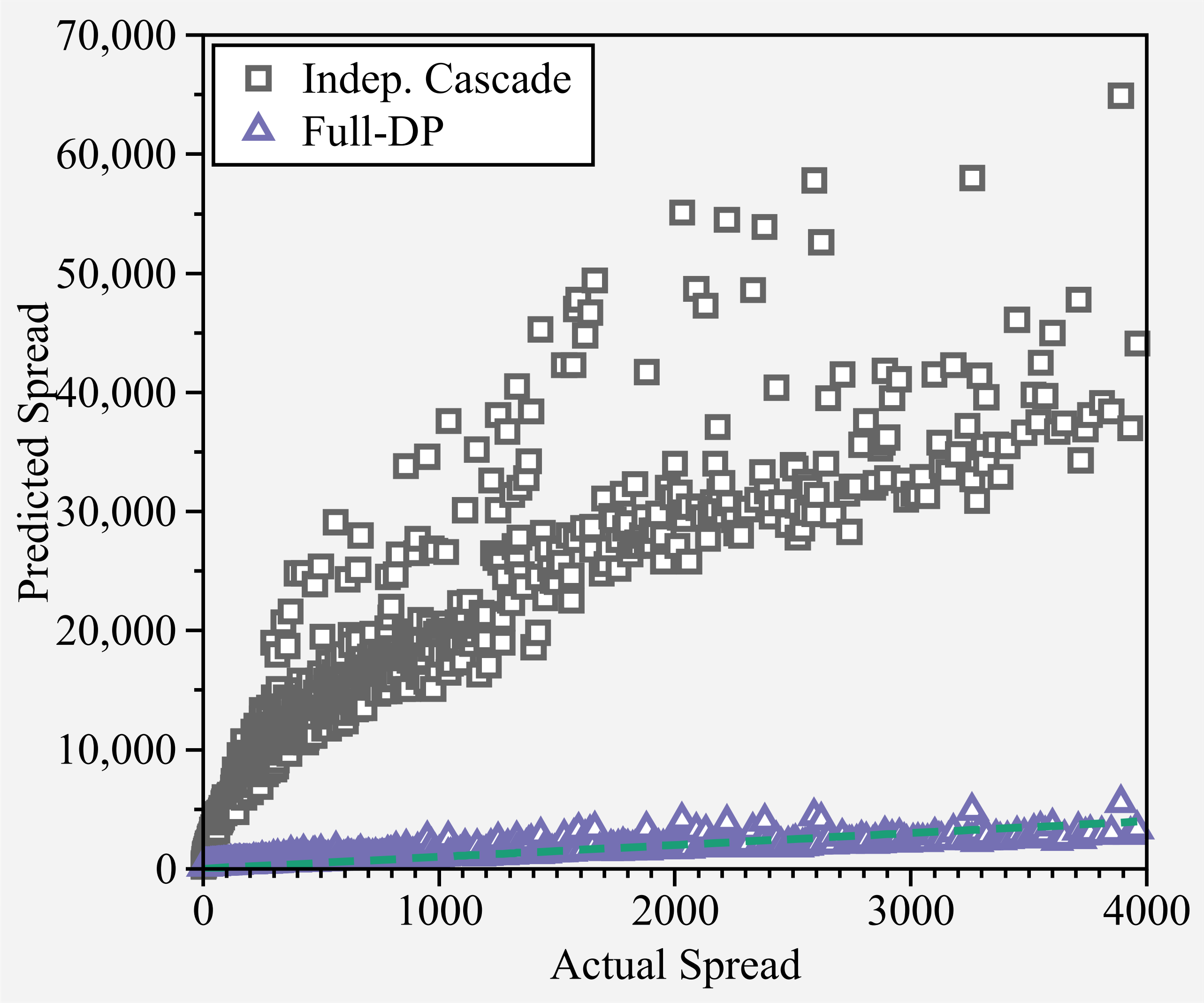
Results - Epinions



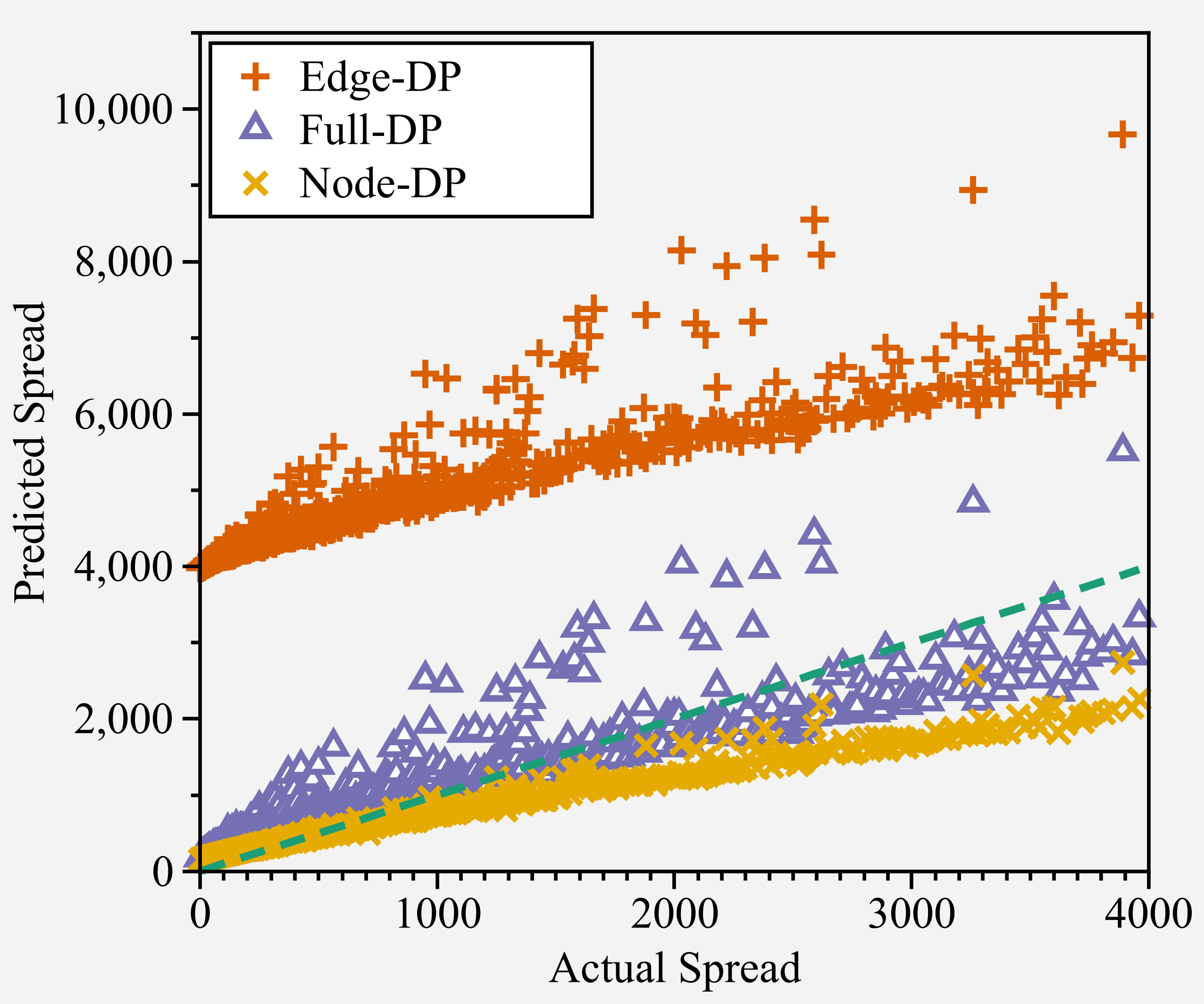
RESULTS - EPINIONS



RESULTS - FLIXSTER



RESULTS - FLIXSTER



All Three Measures

- Directed Propagation matches real-world spreads
- … does best when trained with minimal real-world data
- … and runs quickly (more later)

IDENTIFYING MAXIMIZING SEEDS

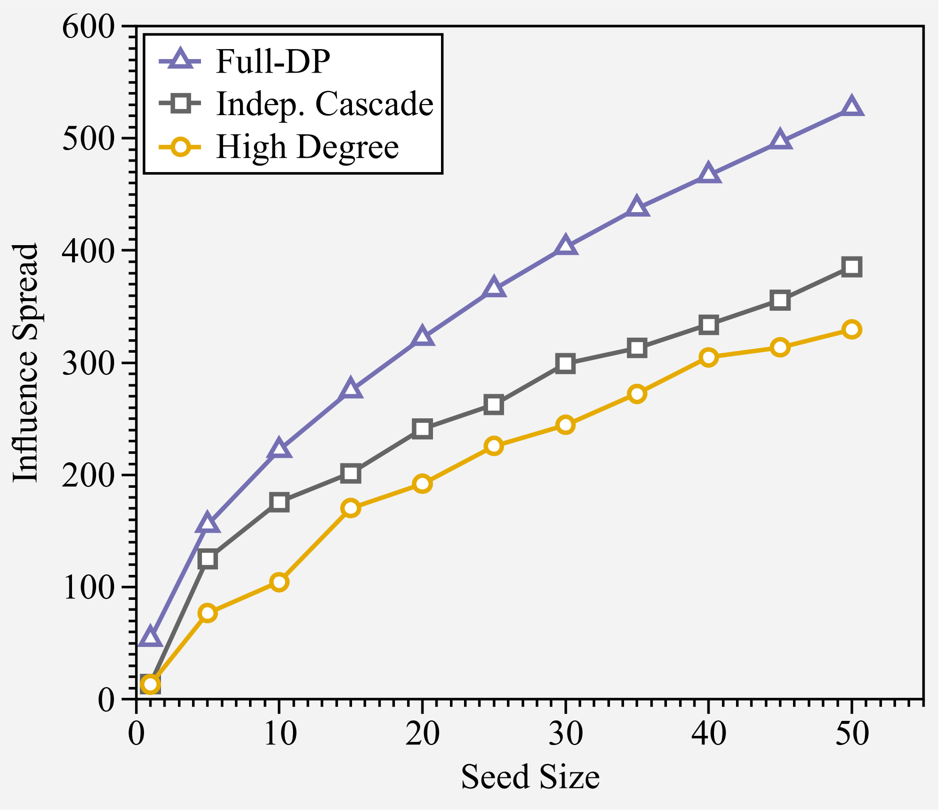
- Full k-seed influence maximization is NP-Hard
  - Greedy algorithm widely used (Kempe, 2003)
  - CELF gives same set; more efficient (Leskovec, 2007)
- Algorithms tested
  - IC (Epinions only)
  - High Degree*
  - Full-DP
  - Edge-DP

*High Degree selects nodes solely based on degree – does not require CELF runs.

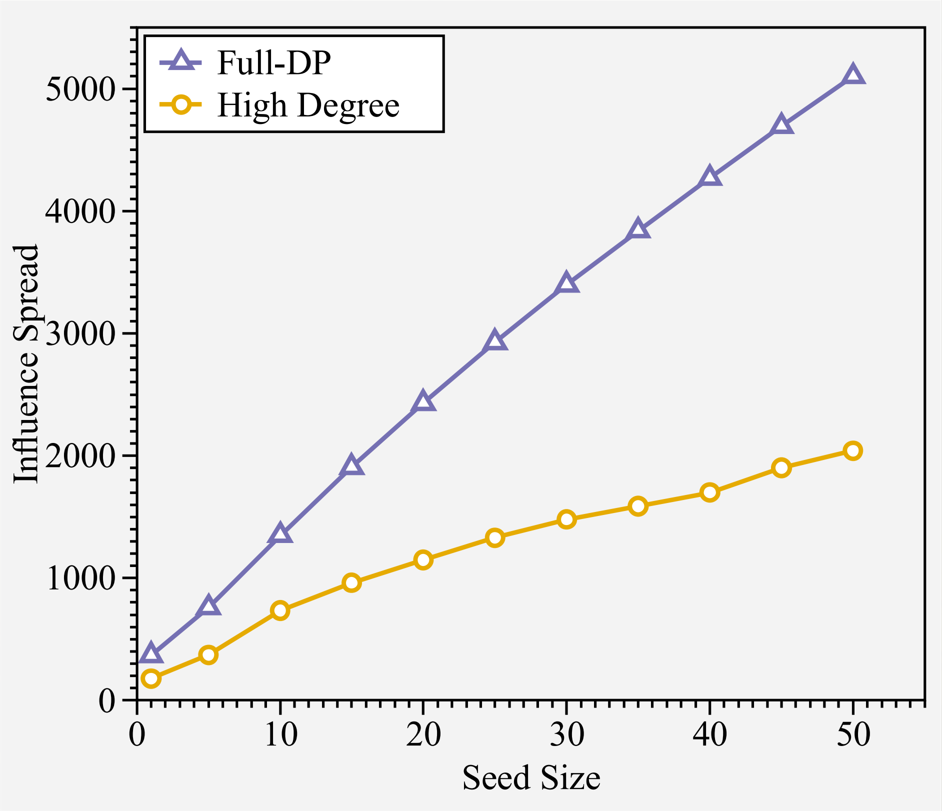
EXPERIMENT

- For given diffusion method (IC, Full-DP, Edge-DP), compute 50 most influential nodes using CELF
- For common comparison, compute diffusion spread for those seeds using Full-DP
  - … as Full-DP was the most accurate to real-world spreads

Results
Influence Maximization



Epinions



Flixster

*Edge-DP not shown as results are almost perfectly coincident with Full-DP

Results
Influence Maximization

- Overlap between identified seed sets
  - Full-DP taken as gold standard

Epinions

	10	20	30	40	50
High Degree	3	6	10	16	18
Edge-DP	10	18	29	36	45
Indep. Cascade	6	9	14	18	24

Flixster

	10	20	30	40	50
High Degree	0	1	3	4	5
Edge-DP	10	20	30	40	49

SEED FEATURES

- How do different methods' selected seeds differ?
  - *Community Detection* – Full-DP and IC chose seeds in separate communities more than Max Degree
  - *Average Degree* – Full-DP chose seeds further apart than IC which chose further apart than Max Degree
  - *Node Degree* – Full-DP chose lower degree nodes than IC which chose lower degree nodes than Max Degree
    - Full-DP and IC chose nodes well above average degree
- Balance between higher degree and distance between seeds

Compute Resources

- Full-DP (Flixster maximization)
  - Initial computation for each node as seed: 12 hours on 60 compute nodes
  - CELF identification of 50 top nodes: 16 minutes on workstation
  - Average propagation: 4 seconds
    - Contrast IC with 10,000 MC simulations: 6 minutes

CONCLUSION

- Directed Propagation
  1. More accurate to real-world data
  2. Easily learned parameters
     - Can identify high-influence nodes without learned params
  3. Computationally efficient

THANKS

- tong@sandia.gov
- jdwendt@sandia.gov

The authors are grateful to Cristopher Moore, David Zage, and Rich Field