Exceptional service in the national interest

Sandia National Laboratories

SAND2016-8074C

ON DATA COLLECTION, GRAPH CONSTRUCTION, AND SAMPLING IN TWITTER

Jeremy D. Wendt, Randy Wells, Richard V. Field, Jr., Sucheta Soundarajan

SYRACUSE UNIVERSITY

International Symposium on Foundations and Applications of Big Data Analytics (FAB) 2016

U.S. DEPARTMENT OF ENERGY

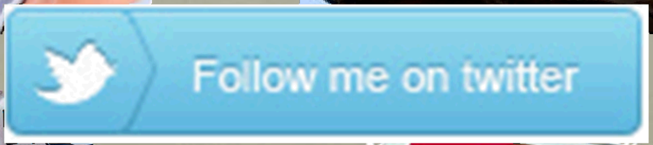
NNSA National Nuclear Security Administration

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.

OVERVIEW

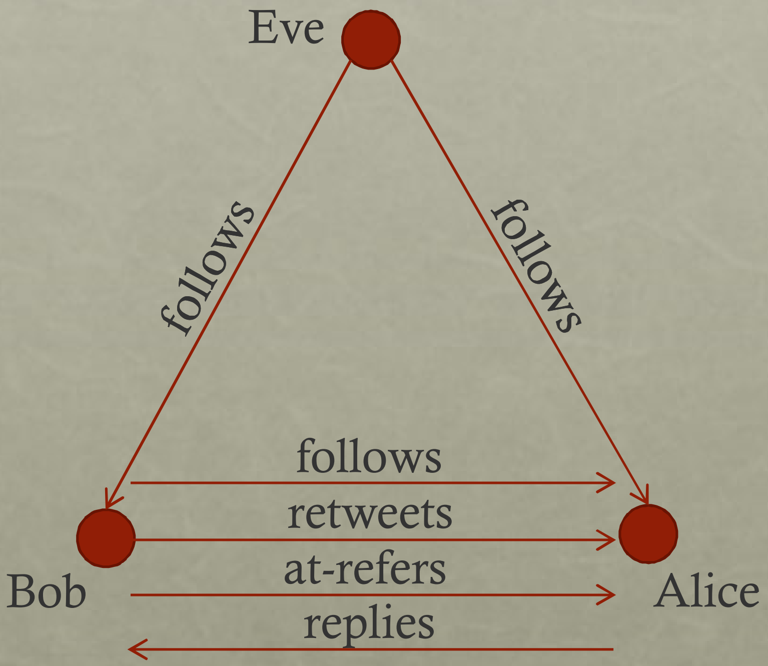
- We present several problems
- We propose some solutions, metrics, and models
  - Fewer than the number of problems

Why Twitter Graphs?



We need to better understand how information flows through Twitter's network

TWITTER'S GRAPH



- Starting state:
  - Bob follows Alice
  - Eve follows Bob and Alice
- Series of tweets:
  - *Alice*: This is a great article: http://some.url
  - *Bob*: (retweeting Alice) This is a great article: http://some.url
  - *Bob*: @Alice, that article was great
  - *Alice*: (reply to Bob) Then you'll love this one: http://other.url

CAN'T GET ALL OF IT

- Twitter allows anyone free access to their data
  - Severely rate limited
  - Different rates for different query types
- 305M active users*
  - >580 years to get all of those
- Therefore, we must sample
  - How does this affect biasing?

5

* http://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/

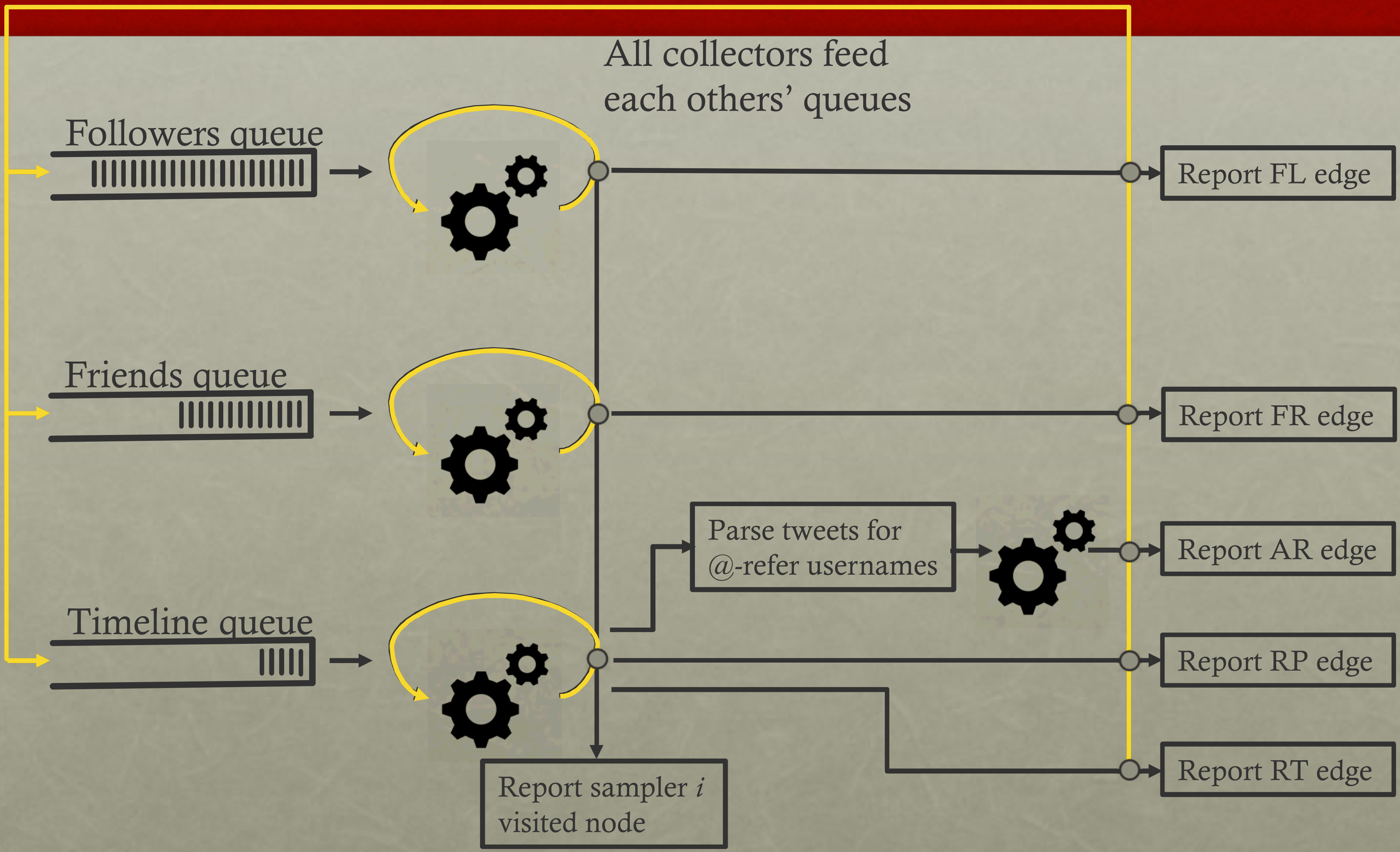
SOME PREVIOUS WORK

- Considerable work on sampling techniques in single-edge-type graphs
  - e.g., (Leskovec, Faloutsos 2006), (Maiya, Berger-Wolf, 2010)
- Sampling introduces bias
  - Random walk finds more high-degree nodes (Lovász, 1993)
  - Bias can be exploited if understood (Maiya, Berger-Wolf, 2011)
- Sampling Twitter
  - Focus on specific edge type (Avrachenkov, et al., 2014)
- Ours appears to be first work focused on sampling in multiple-edge-type networks

GETTING TWITTER

- Twitter provides free access to their data via the Twitter API
  - Different queries for different edge types
  - Rate limits vary for different queries
  - Multiple queries to get full information for a single user
- **Problem 1:** How do you sample different requests at different rates?
  - We propose separate queues for each
- **Problem 2:** How do you keep the queues from sampling different parts of the graph?
  - We propose shared-fed queues

Our Twitter Collector



All collectors feed each others' queues

Followers queue

Friends queue

Timeline queue

Parse tweets for @-refer usernames

Report FL edge

Report FR edge

Report AR edge

Report RP edge

Report RT edge

Report sampler *i* visited node

Collection Results 1

Number of Requests

ID	Duration (days)	Friend	Follower	Timeline
1	7	7,773	7,259	139,540
2	9	8,690	9,002	168,822
3	7	6,511	6,670	118,682

COLLECTION RESULTS 2

Number of Users

ID	Duration (days)	Friend	Follower	Timeline
1	7	4,435	118	13,573
2	9	4,797	878	11,319
3	7	3,780	166	10,050
Ave	Req. per	1.8	37.3	12.3

Collection Results 3

Users with Zero Results

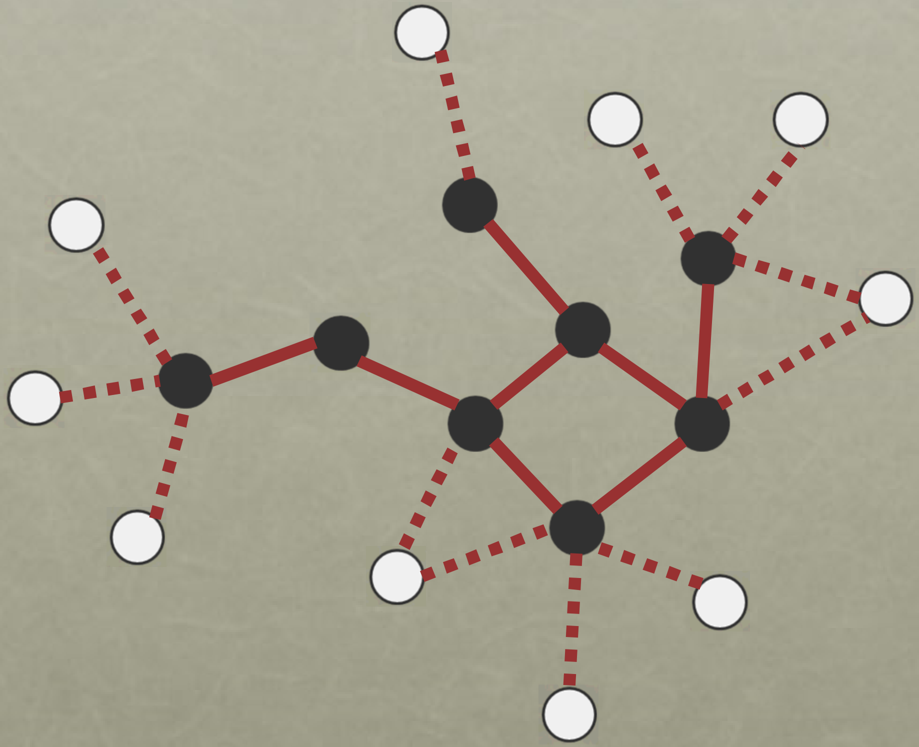
ID	Duration (days)	Friend	Follower	Timeline
1	7	37%	11%	6%
2	9	24%	33%	5%
3	7	44%	22%	20%
Ave	0-queries (hours)	35.6	2.7	1.5

COLLECTION PROBLEMS

- **Problem 3:** Can we avoid more of those zero-hits queries?
  - There may be indications between collectors' results that push away from zero-hits

- **Problem 4:** How would avoiding the highest degree follower nodes affect biasing?

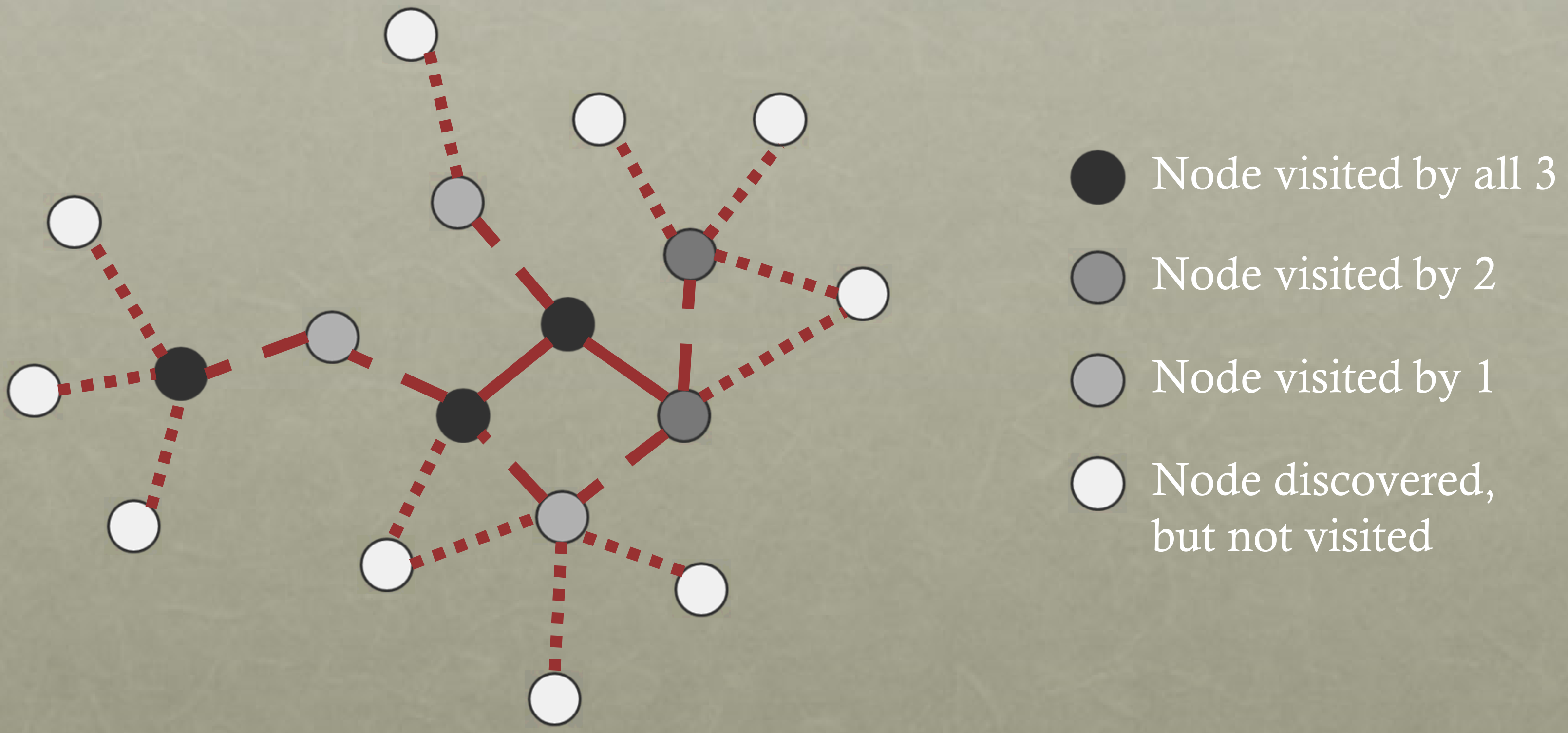
TRADITIONAL SAMPLING



● Node visited

○ Node discovered, but not visited

SEMANTIC SAMPLING



● Node visited by all 3

● Node visited by 2

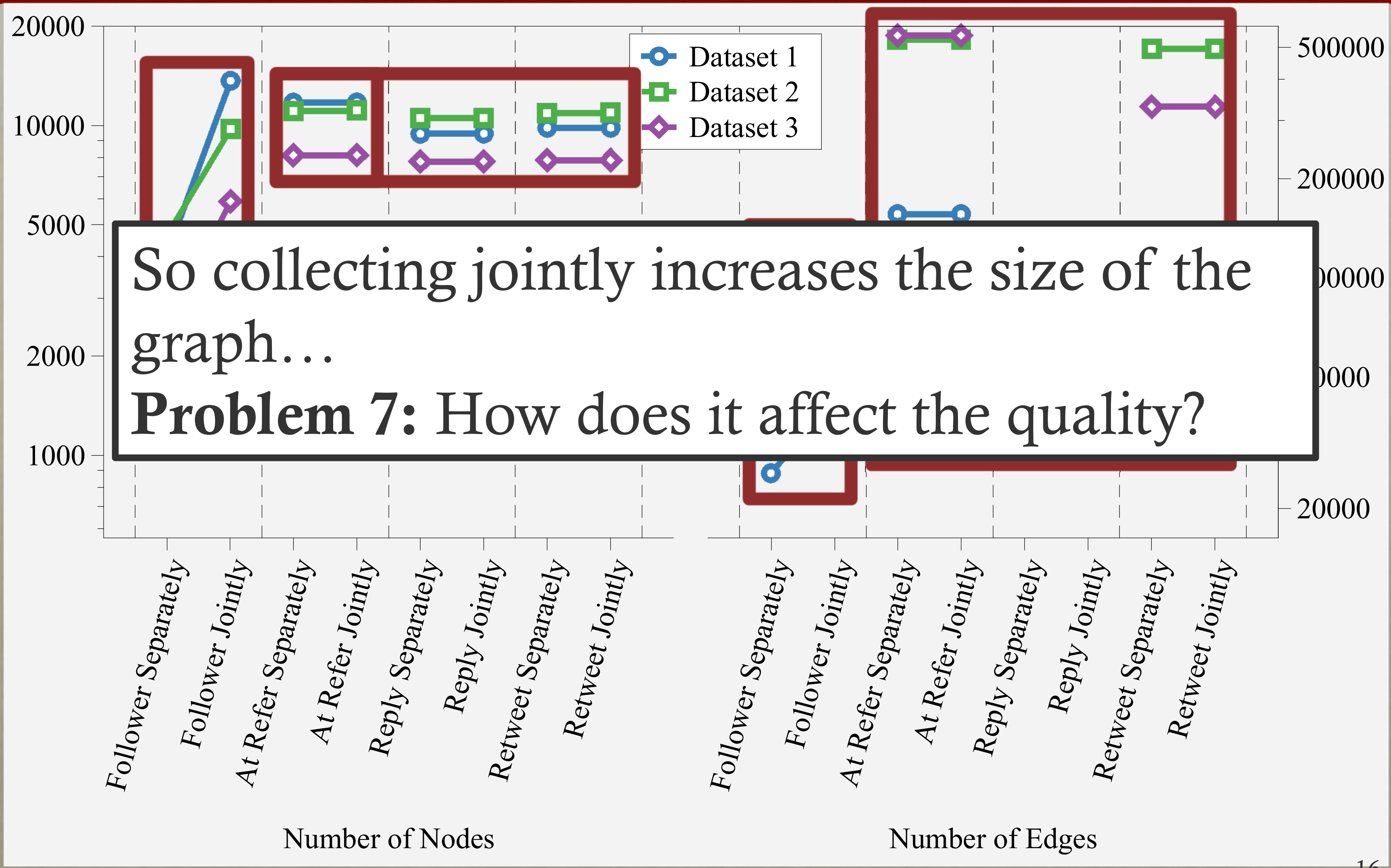
● Node visited by 1

○ Node discovered, but not visited

Forming a Graph

- **Problem 5:** How do we define "visited" with multiple collectors?
  - We propose visited by any
- **Problem 6:** Which edges are allowed in the graph?
  - We propose two options
- *Collecting separately* requires the edge-type sampler to visit both ends for an edge of that type to be included
- *Collecting jointly* request some edge-type sampler to visit both ends for an edge of any type to be included

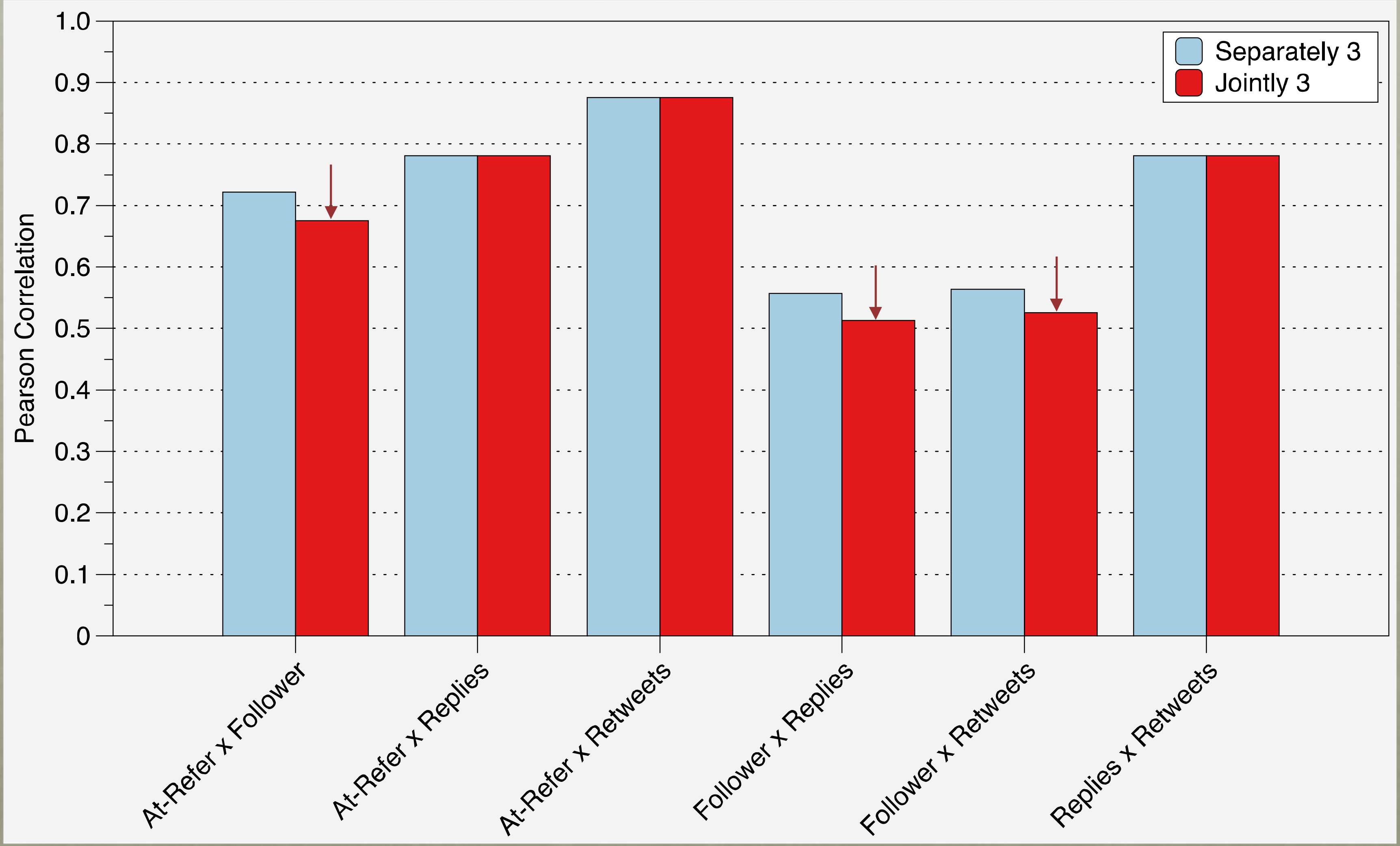
Separately vs. Jointly



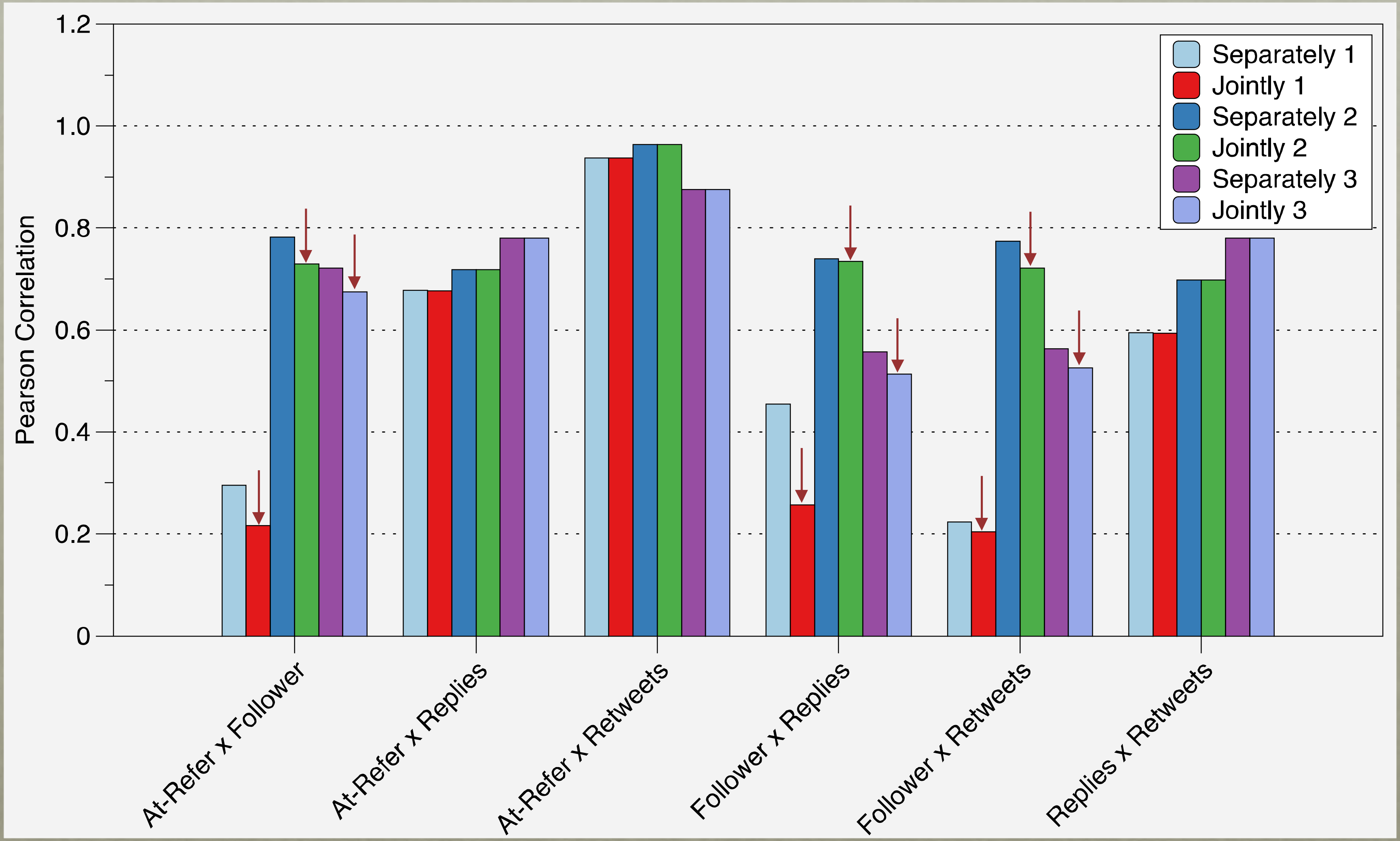
So collecting jointly increases the size of the graph…

Problem 7: How does it affect the quality?

Graph Metrics



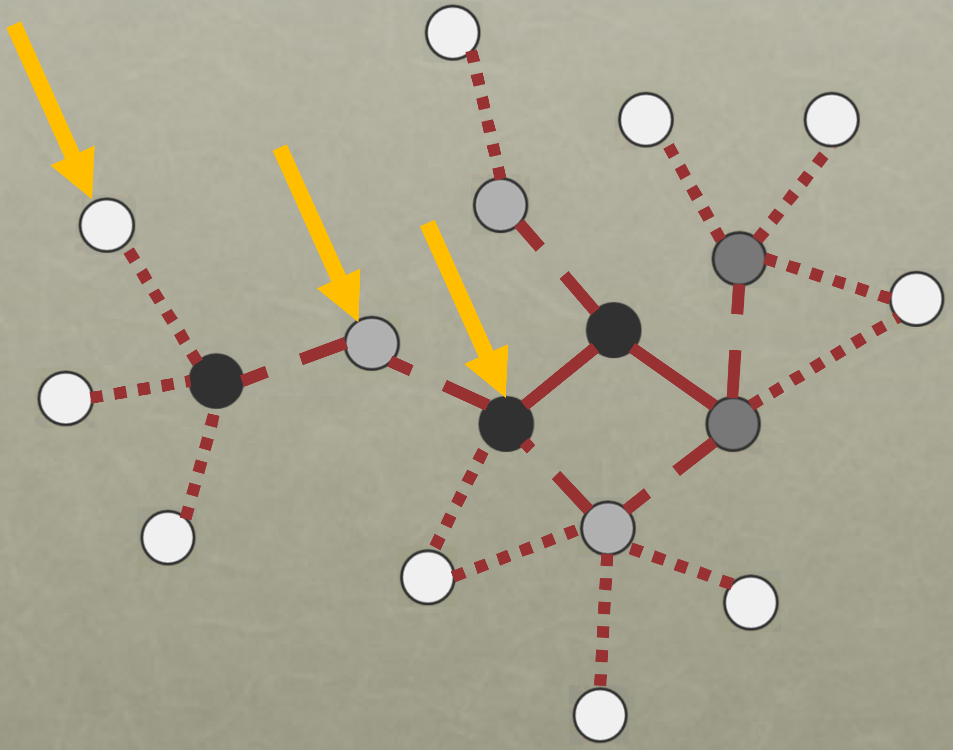
Graph Metrics



SUBSAMPLING

- Traditional graph sampling samples a graph using different techniques and analyzes how different techniques affect metrics
- We attempted this with our semantic graph…

SUBSAMPLING



● Node visited by all 3

● Node visited by 2

● Node visited by 1

○ Node discovered, but not visited

SAMPLING ANALYSIS

Friends

ID	Successes (1 hour)	Failures (1 hour)
1	60	816
2	60	548
3	60	723

Followers

ID	Successes (1 hour)	Failures (1 hour)
1	60	4,589
2	60	28,189
3	60	644

Timeline

ID	Successes (1 hour)	Failures (1 hour)
1	1,200	1,052
2	1,200	1,141
3	1,200	625

- **Problem 8:** We hit the edge of our collect more than getting good results. Why? How can we avoid it?

Why Failures?

- Asynchronous queries mean you get responses from one collector before another
  - The to-visit queue is thus in a different order
- Our graphs are relatively small … even though collected over many days
- We propose a model on why we are so close to the collection “edge” even at the seed – curse of dimensionality
  - Different collectors add more dimensions to the data
  - Increased dimensionality → decreased density
  - Sparsest collector (followers) makes a narrow dimension

SEMANTIC GRAPHS SPECIFIC?

- Traditional graph sampling has not reported this “edge failure” phenomenon
  - However, they subsampled against the graph itself
    - Hitting a leaf can mean degree 1 in original data or a collection edge
  - We sampled against original collected data
    - Thus, we could differentiate between true leaves and collection-caused leaves
- **Problem 9:** How much does this occur in traditional graphs?

OPEN PROBLEMS

- How do you sample different requests at different rates?
- How do you keep the queues from sampling different parts of the graph?
- Can we avoid more of those zero-hits queries?
- Can we avoid the highest degree follower nodes with minimal biasing?
- How do we define “visited” with multiple collectors?
- Which edges are allowed in the graph?
- How does collecting joinly vs. collecting separately affect graph quality?
- We hit the edge of our collect more than getting good results. Why? How can we avoid it?
- How much does edge-hitting exist in traditional graphs?

WE NEED ANSWERS!

THANKS

- jdwendt@sandia.gov
- rwells@sandia.gov
- rvfield@sandia.gov
- susounda@syr.edu