

Benchmarking API Costs of Network Sampling Strategies

Michele Coscia & Luca Rossi
ITU København

August 27th, 2019

IT UNIVERSITY OF COPENHAGEN

(i) Why Network Sampling?

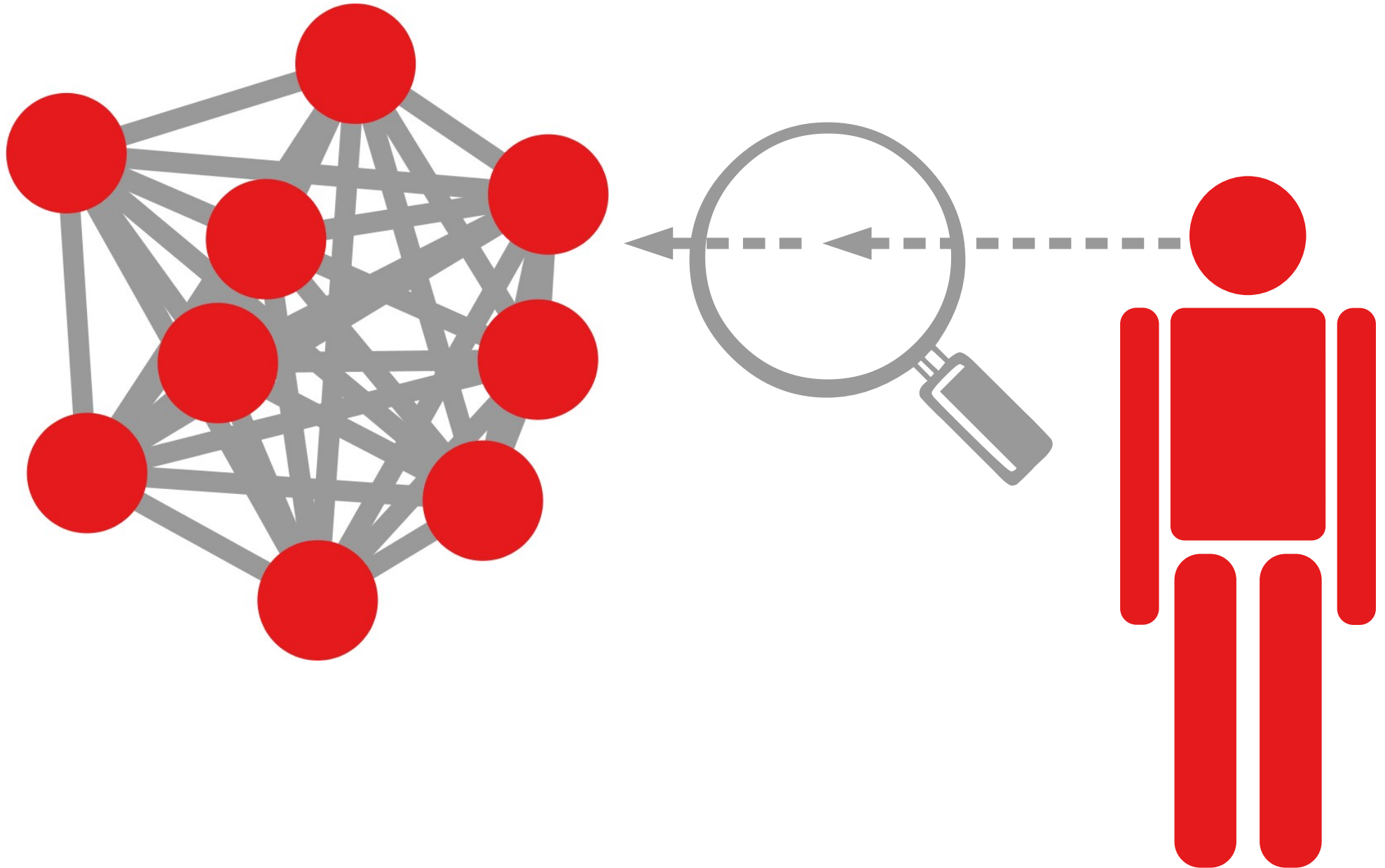
(i) Why Network Sampling?

(ii) Are there understudied real world obstacles that should make us reconsider how we choose the best sampling strategy?

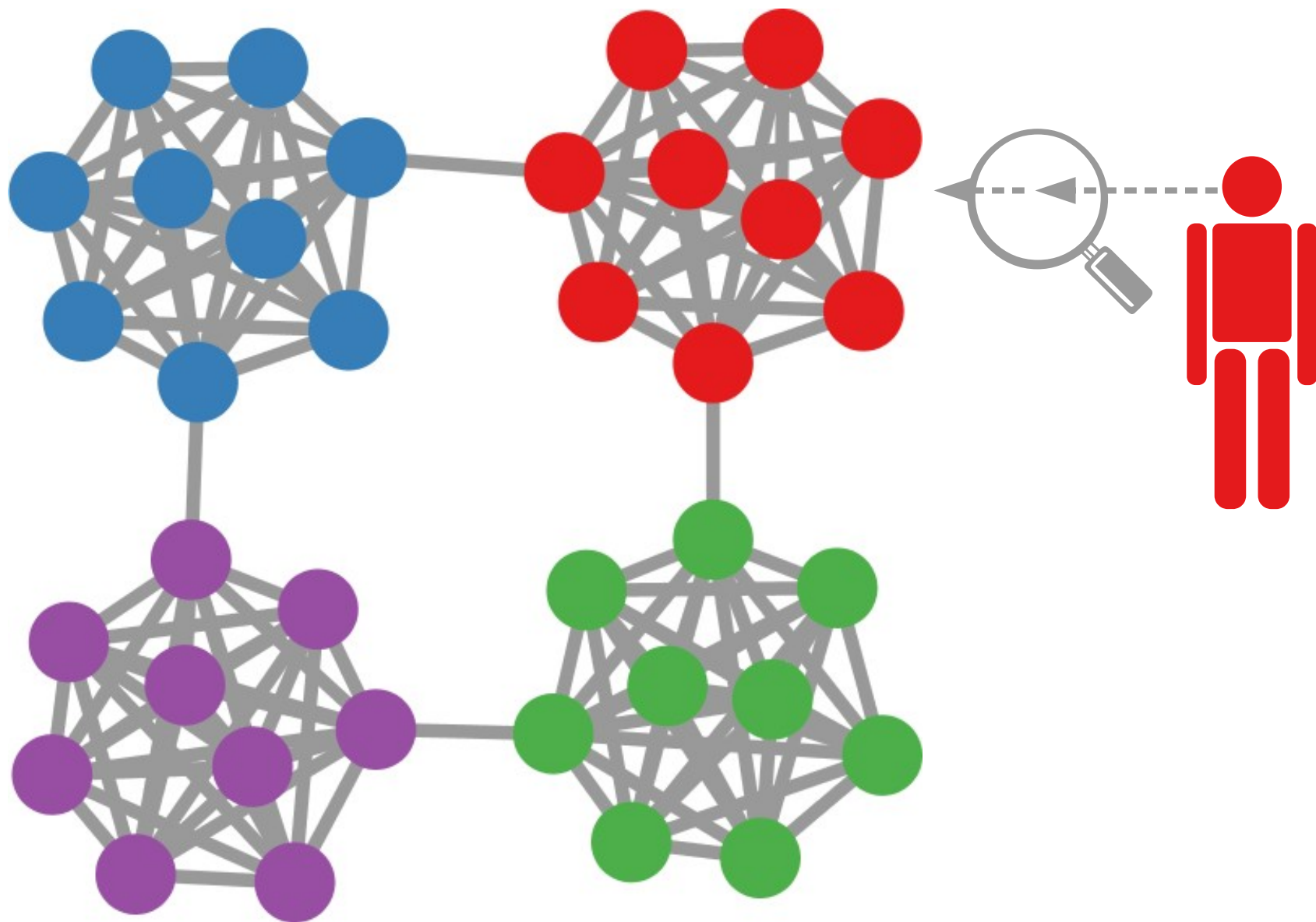
(i) Why Network Sampling?

The Observation Problem

The Observation Problem



The Observation Problem







~330M monthly users



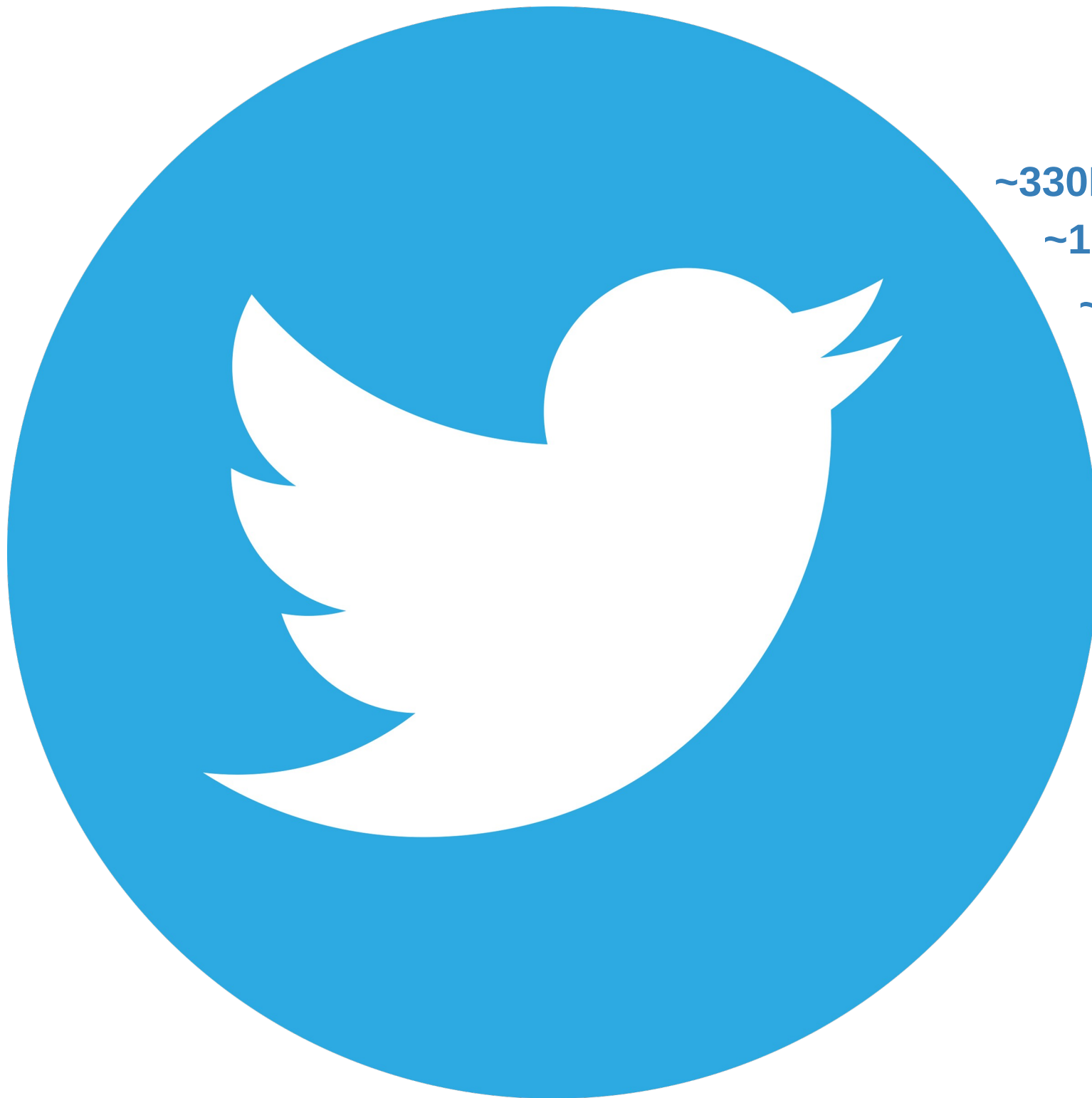
~330M monthly users
~1.1m per user



~330M monthly users

~1.1m per user

~21.78B seconds

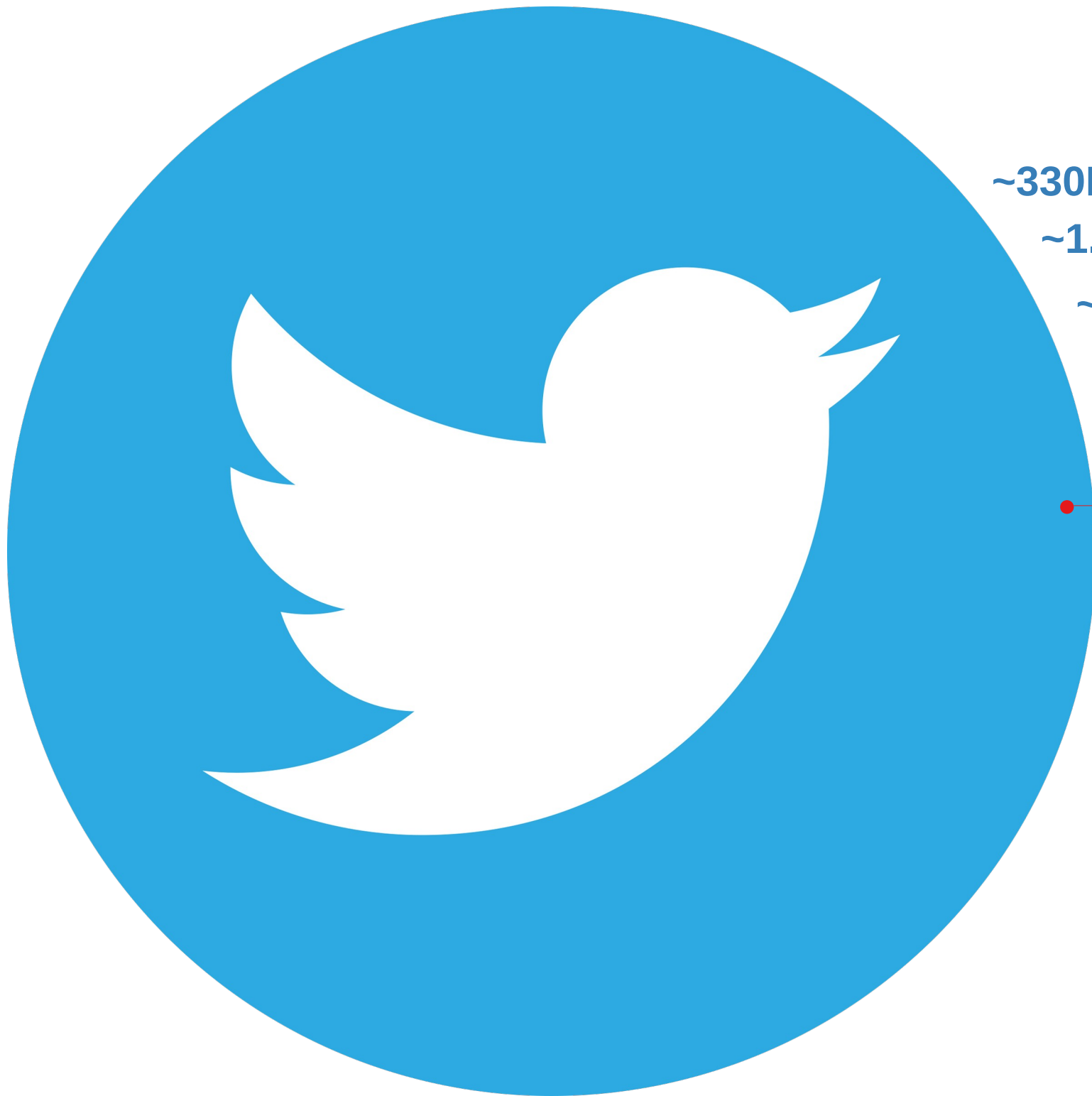


~330M monthly users

~1.1m per user

~21.78B seconds

~690 years



~330M monthly users

~1.1m per user

~21.78B seconds

~690 years

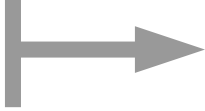


**1 year of
crawling**



Network Exploration

- BFS, DFS
- Random Walks
- Snowball
- Forest Fire

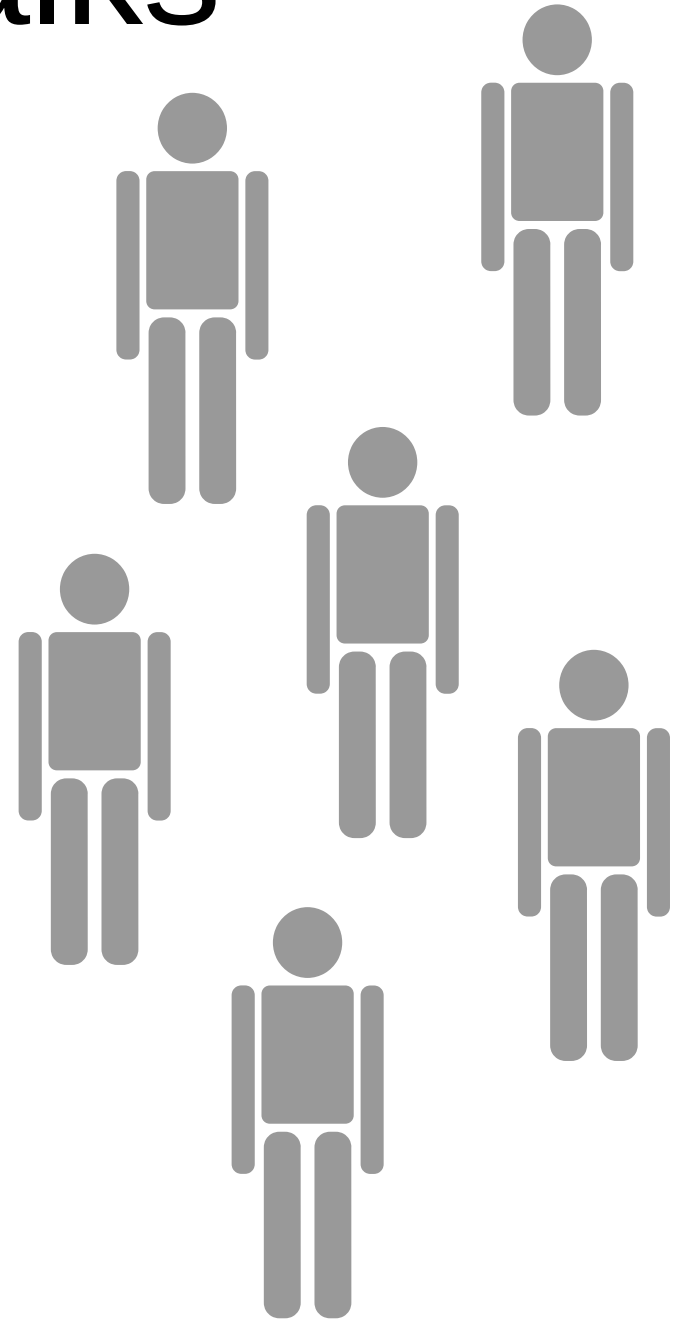
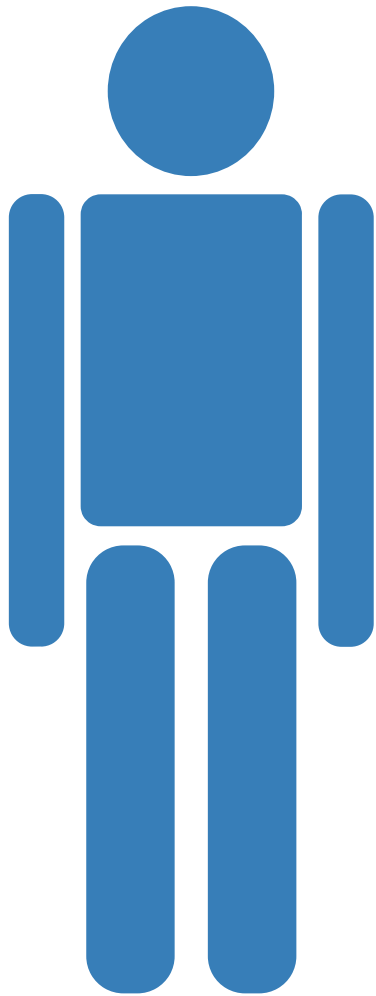
Network Exploration

- BFS, DFS  Full exploration as the objective
- Random Walks
- Snowball
- Forest Fire

Network Exploration

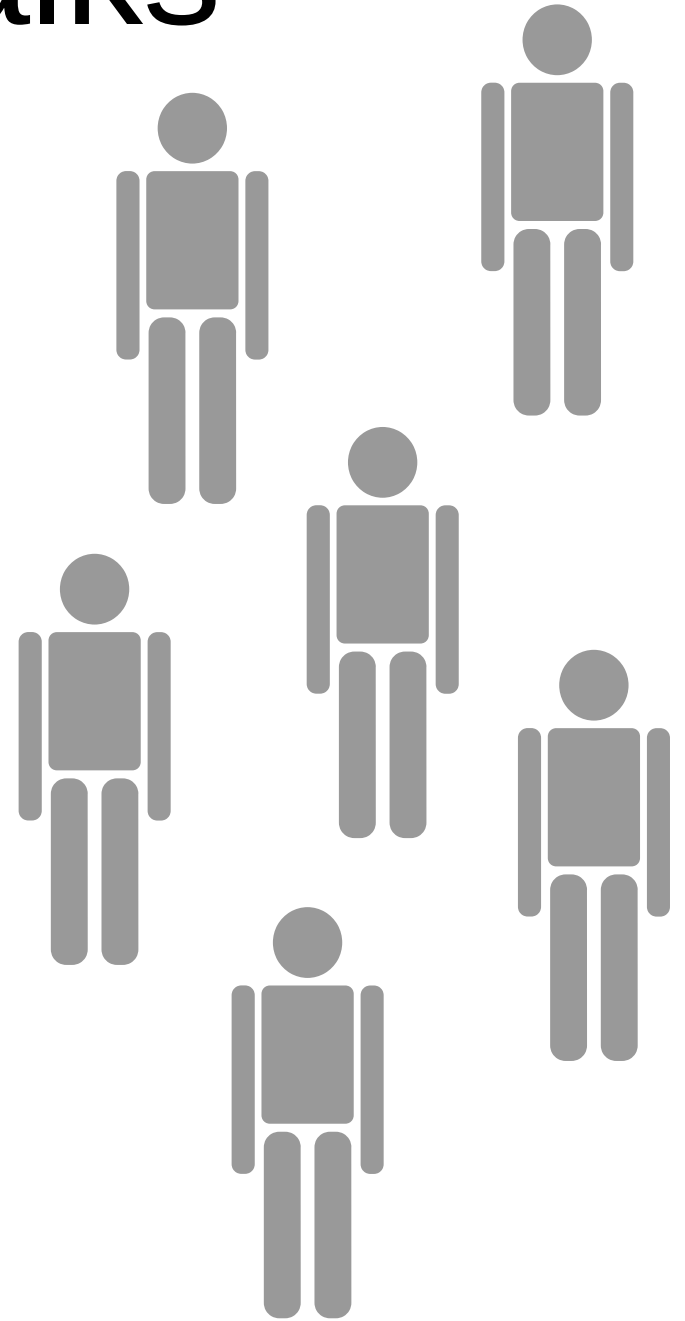
- BFS, DFS  Full exploration as the objective
 - Random Walks
 - Snowball
 - Forest Fire
-  Preventing bias from samples

Random Walks



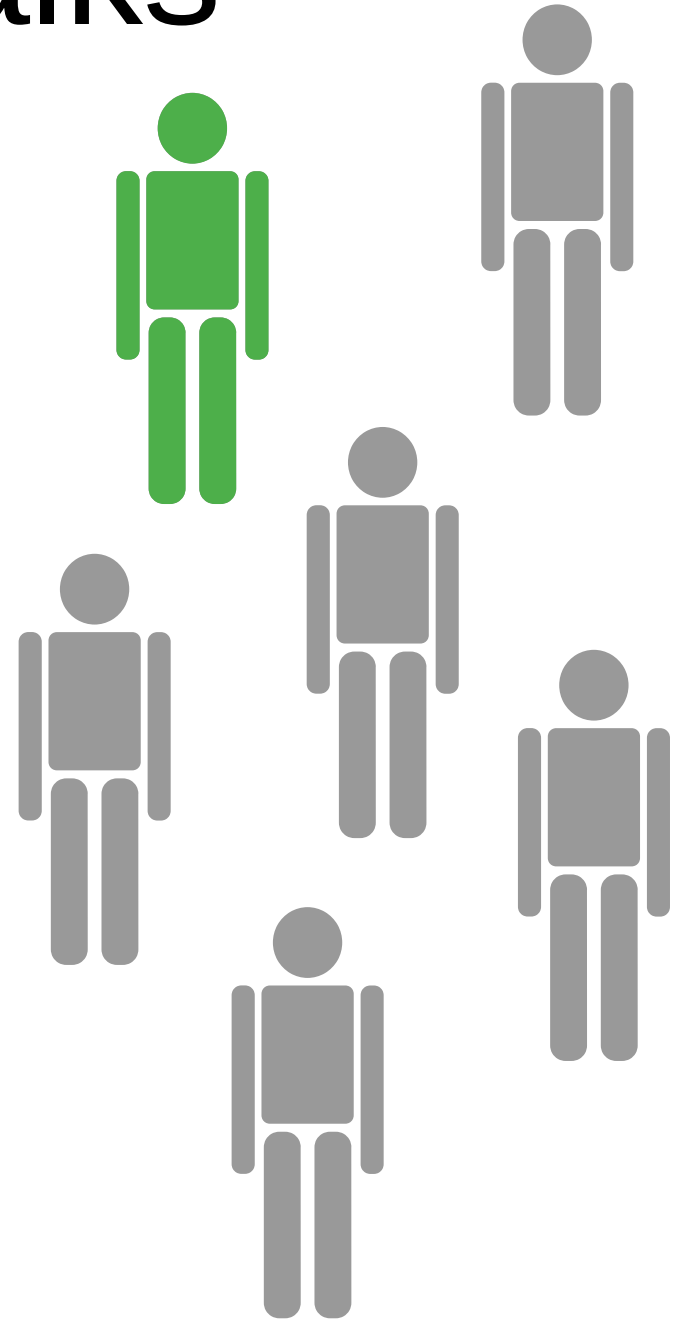
Random Walks

**Name 1
of your
friends**

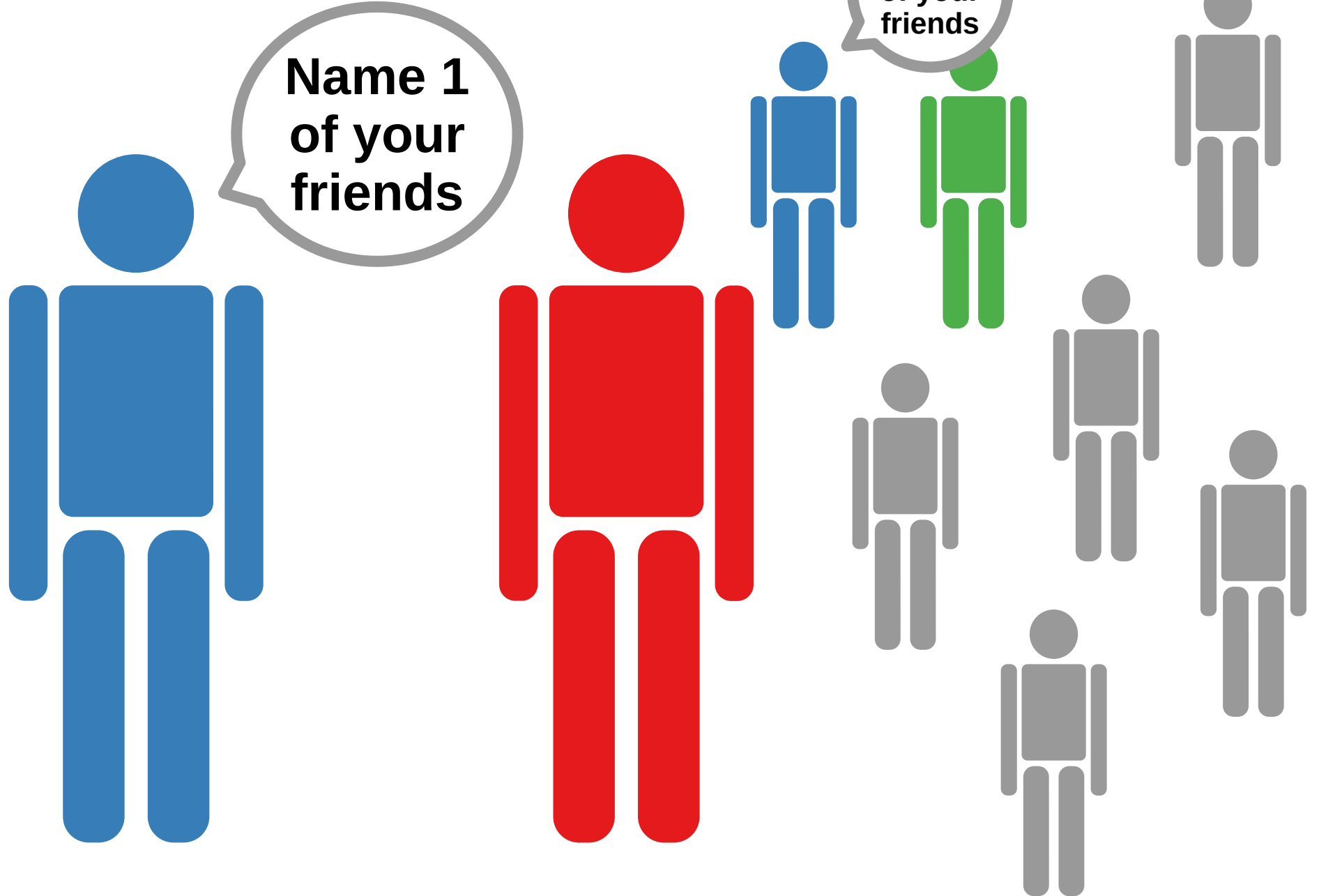


Random Walks

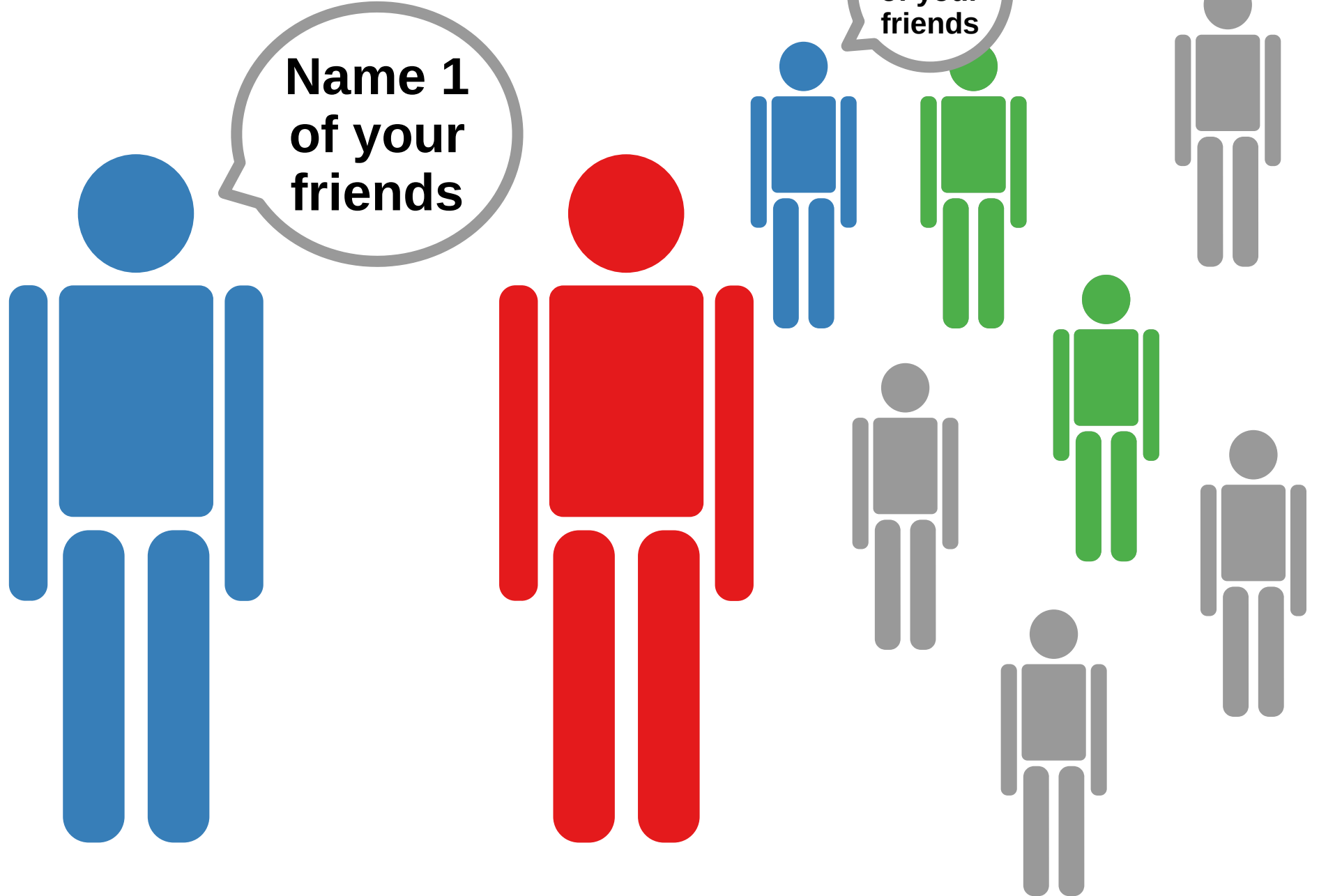
**Name 1
of your
friends**



Random Walk



Random Walk



Degree Bias

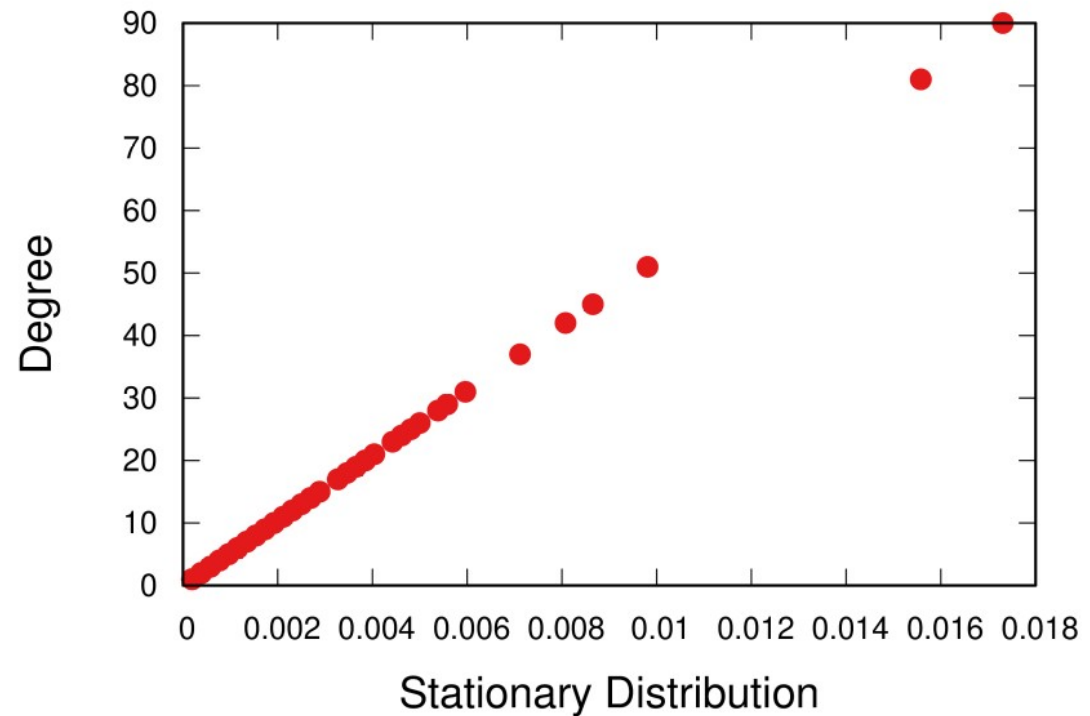
Degree Bias

- Stationary distr π

[illegible]

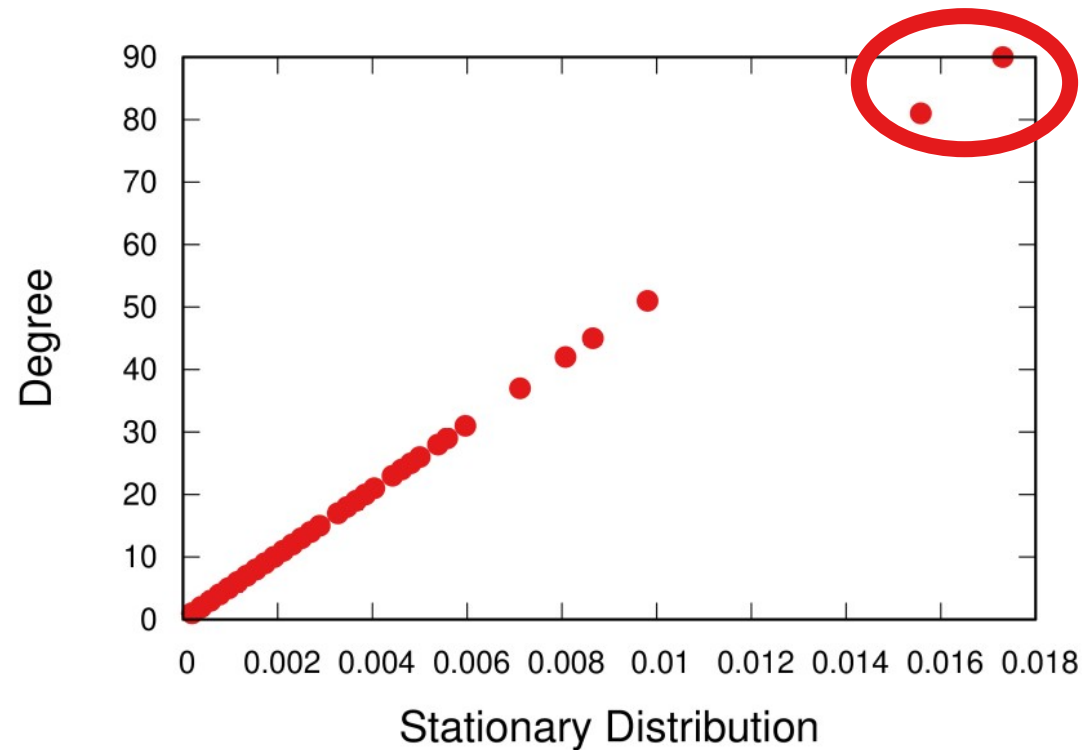
Degree Bias

- Stationary distr
 π
- $\pi = \text{degree}$

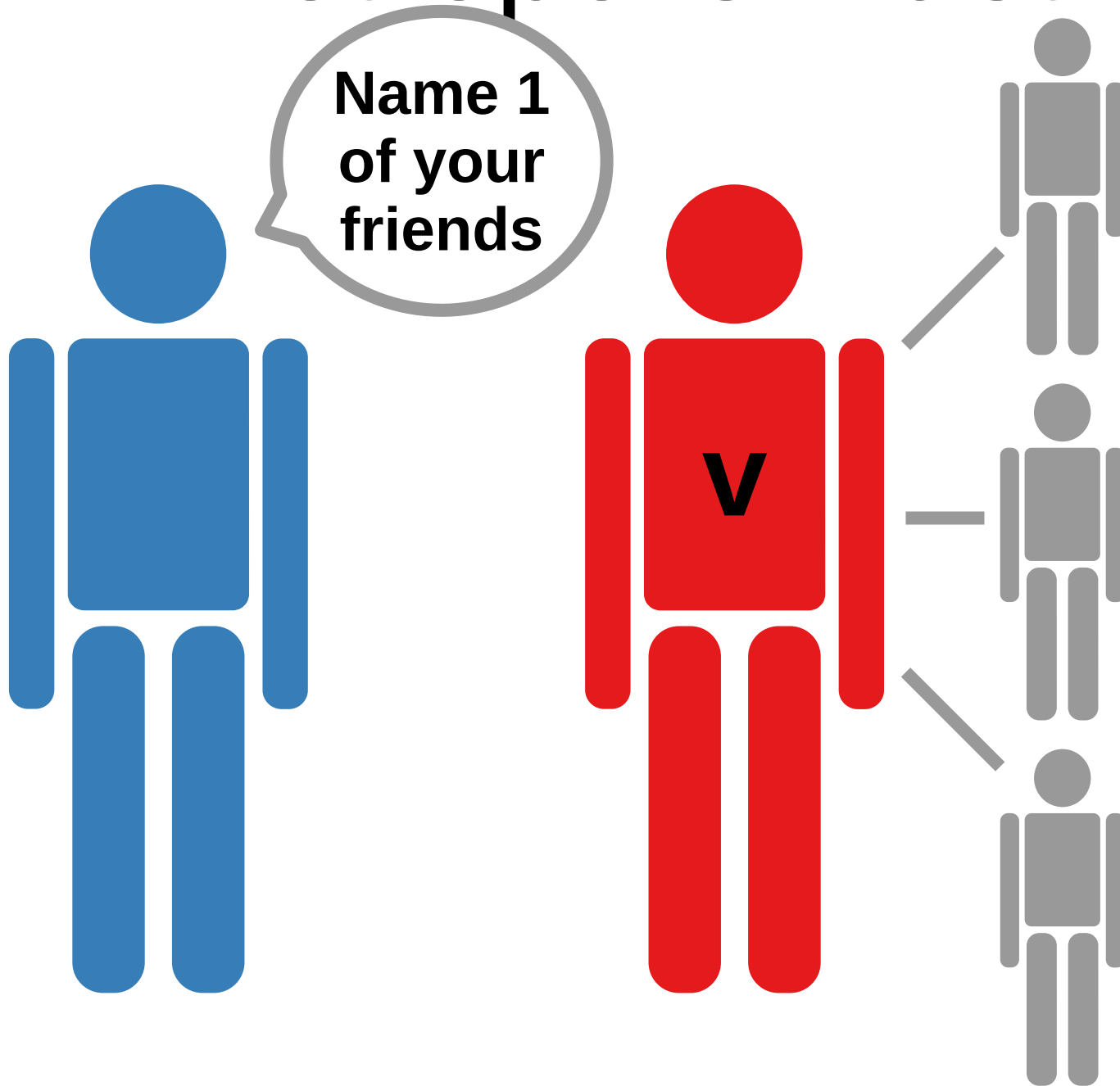


Degree Bias

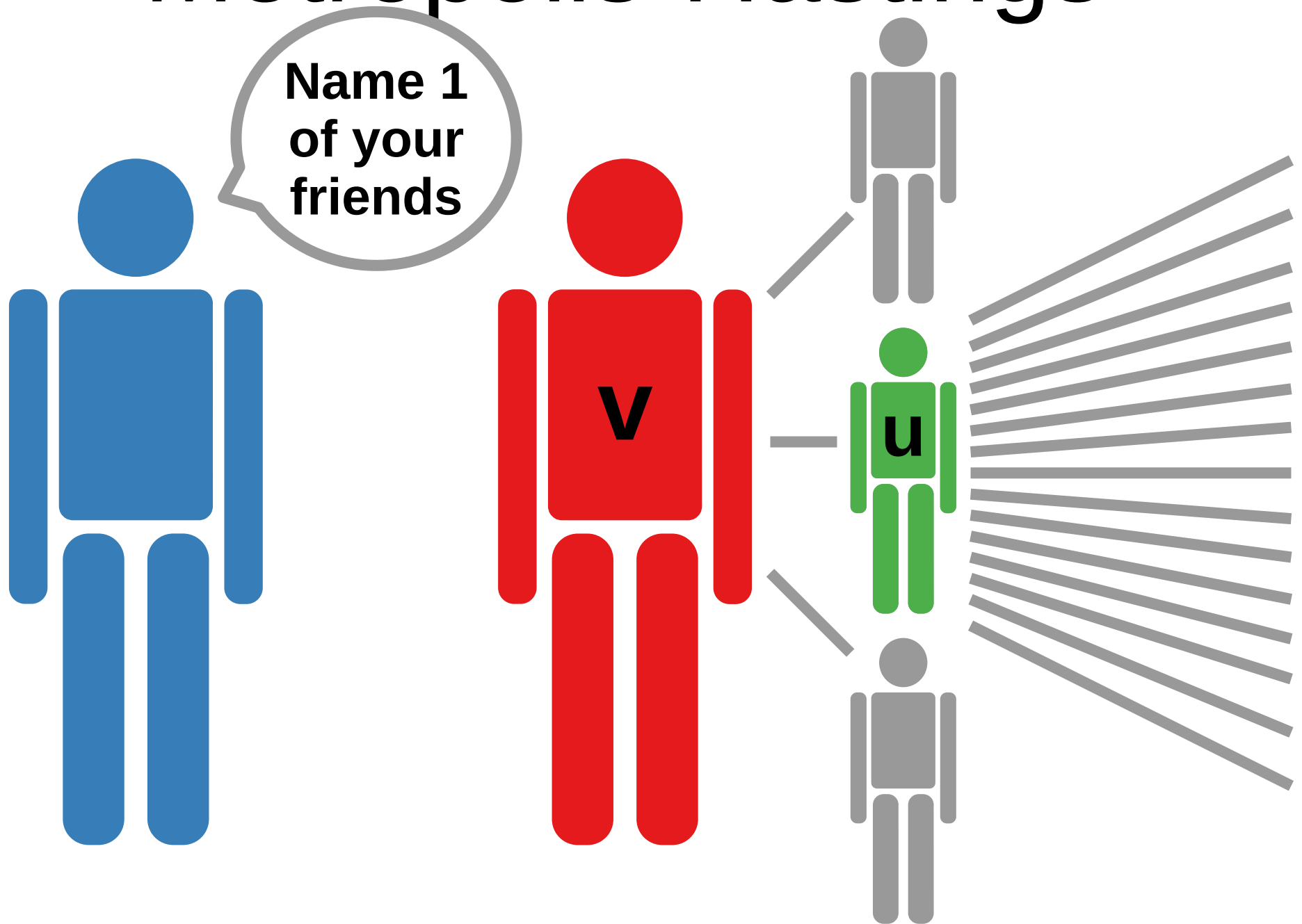
- Stationary distr π
- $\pi = \text{degree}$
- Oversampled hubs!



Metropolis-Hastings



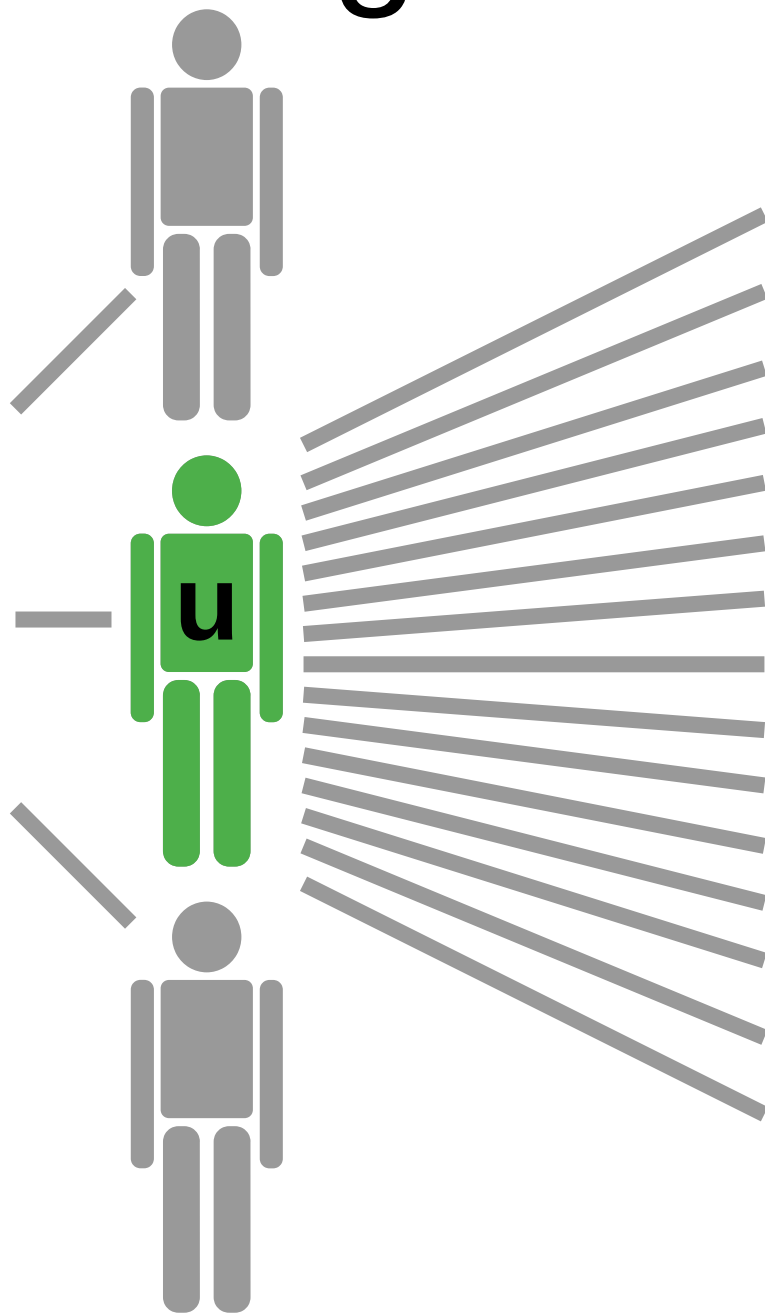
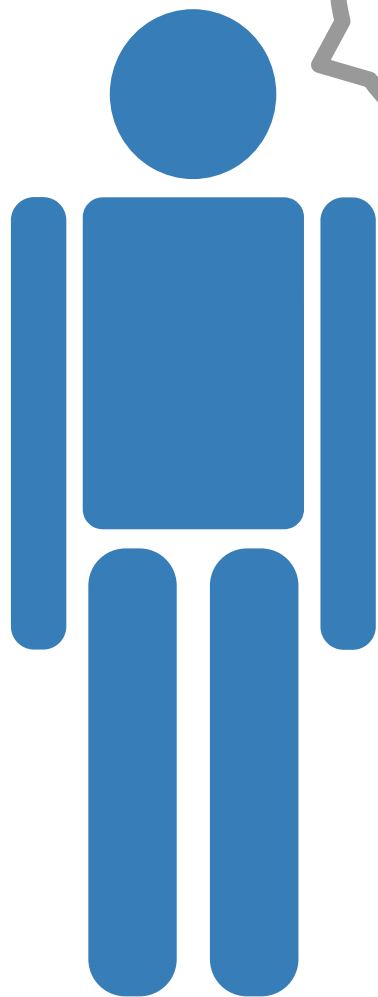
Metropolis-Hastings



Metropolis-Hastings

$$p \sim k_v / k_u$$

Name 1
of your
friends



Metropolis-Hastings

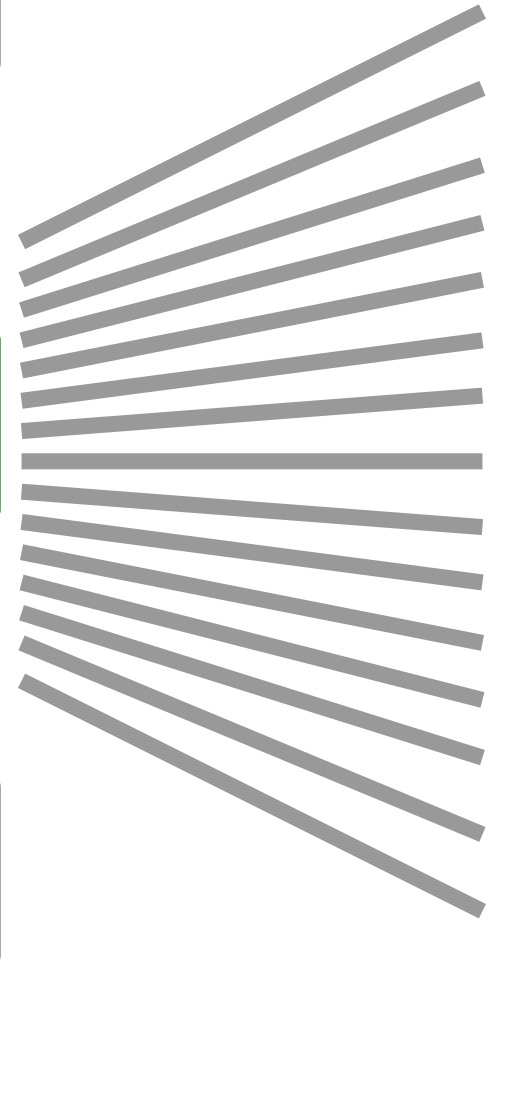
$$p \sim k_v / k_u$$

Name 1
of your
friends

Mmm...
Name
another

v

u



Re-Weighted RW

Re-Weighted RW

- Perform vanilla
RW

Re-Weighted RW

- Perform vanilla RW
- Re-weight property of interest

Re-Weighted RW

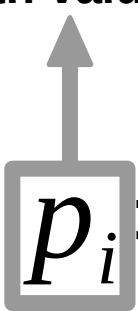
- Perform vanilla RW
- Re-weight property of interest

$$p_i = \frac{\sum_{v \in V_i} i^{-1}}{\sum_{v' \in V} x_{v'}^{-1}}$$

Re-Weighted RW

- Perform vanilla RW
- Re-weight property of interest

p of nodes
with value i


$$p_i = \frac{\sum_{v \in V_i} i^{-1}}{\sum_{v' \in V} x_{v'}^{-1}}$$

Re-Weighted RW

- Perform vanilla RW
- Re-weight property of interest

The diagram illustrates the formula for the probability p_i of a node having value i . It features a central equation with annotations. On the left, a box containing p_i has an upward arrow pointing to the text "p of nodes with value i". The equation itself is
$$p_i = \frac{\sum_{v \in V_i} i^{-1}}{\sum_{v' \in V} x_{v'}^{-1}}$$
 The numerator's summation index $v \in V_i$ is enclosed in a box, with an arrow pointing from this box to the text "Set of nodes with value i" on the right. The denominator's summation index is $v' \in V$.

p of nodes with value i

$$p_i = \frac{\sum_{v \in V_i} i^{-1}}{\sum_{v' \in V} x_{v'}^{-1}}$$

Set of nodes with value i

Re-Weighted RW

- Perform vanilla RW
- Re-weight property of interest

The diagram illustrates the formula for the probability p_i of a node having value i . The formula is:

$$p_i = \frac{\sum_{v \in V_i} i^{-1}}{\sum_{v' \in V} x_{v'}^{-1}}$$

The components of the formula are annotated with arrows and text:

- An arrow points from the boxed p_i to the text "p of nodes with value i".
- An arrow points from the boxed V_i in the numerator to the text "Set of nodes with value i".
- An arrow points from the boxed V in the denominator to the text "Set of nodes in the sample".

Re-Weighted RW

- Perform vanilla RW
- Re-weight property of interest

$$p_i = \frac{\sum_{v \in V_i} i^{-1}}{\sum_{v' \in V} x_{v'}^{-1}}$$

p of nodes with value i

Set of nodes with value i

Set of nodes in the sample

Value for v'

Re-Weighted RW

- Perform vanilla RW
- Re-weight property of interest
- Respondent-Driven Sampling

The diagram illustrates the re-weighting formula for Respondent-Driven Sampling. It shows the relationship between the probability p_i of nodes with value i , the set of nodes with value i , the set of nodes in the sample, and the value for a specific node v' .

$$p_i = \frac{\sum_{v \in V_i} i^{-1}}{\sum_{v' \in V} x_{v'}^{-1}}$$

Annotations in the diagram:

- p_i is boxed, with an arrow pointing to "p of nodes with value i".
- V_i is boxed, with an arrow pointing to "Set of nodes with value i".
- V is boxed, with an arrow pointing to "Set of nodes in the sample".
- $x_{v'}$ is boxed, with an arrow pointing to "Value for v' ".

Re-Weighted RW: Example

Re-Weighted RW: Example

- p of a node having $k=2$?

Re-Weighted RW: Example

- p of a node having $k=2$?
- Observed: 20 over 100 ($p = 0.2$)

Re-Weighted RW: Example

- p of a node having $k=2$?
- Observed: 20 over 100 ($p = 0.2$)
- Other nodes:
 - $k=1$: 50
 - $k=3$: 10
 - $k=4$: 8
 - $k=5$: 7
 - $k=6$: 5

Re-Weighted RW: Example

- p of a node having k=2?
- Observed: 20 over 100 (p = 0.2)

- Other nodes:

- k=1: 50

- k=3: 10

- k=4: 8

- k=5: 7

- k=6: 5

$$p_2 = \frac{20 * 1/2}{(50/1) + (20/2) + (10/3) + (8/4) + (7/5) + (5/6)}$$

Re-Weighted RW: Example

- p of a node having k=2?
- Observed: 20 over 100 (p = 0.2)

- Other nodes:

- k=1: 50

- k=3: 10

- k=4: 8

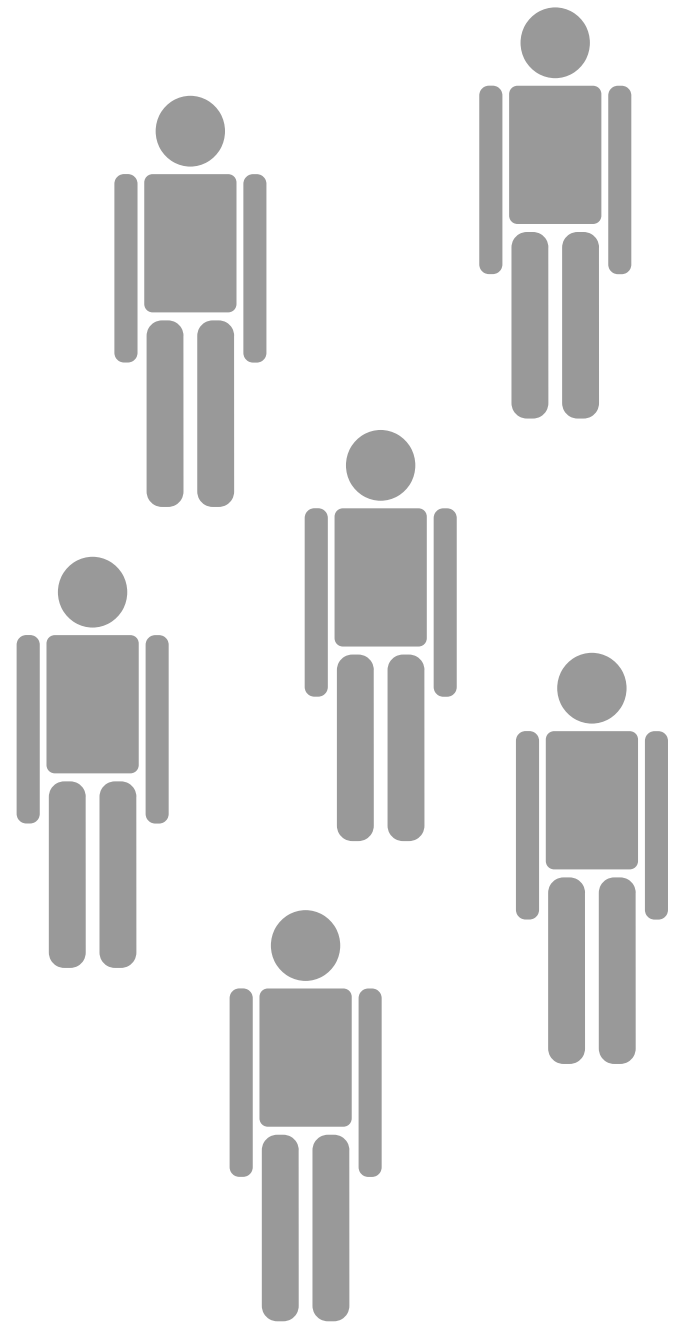
- k=5: 7

- k=6: 5

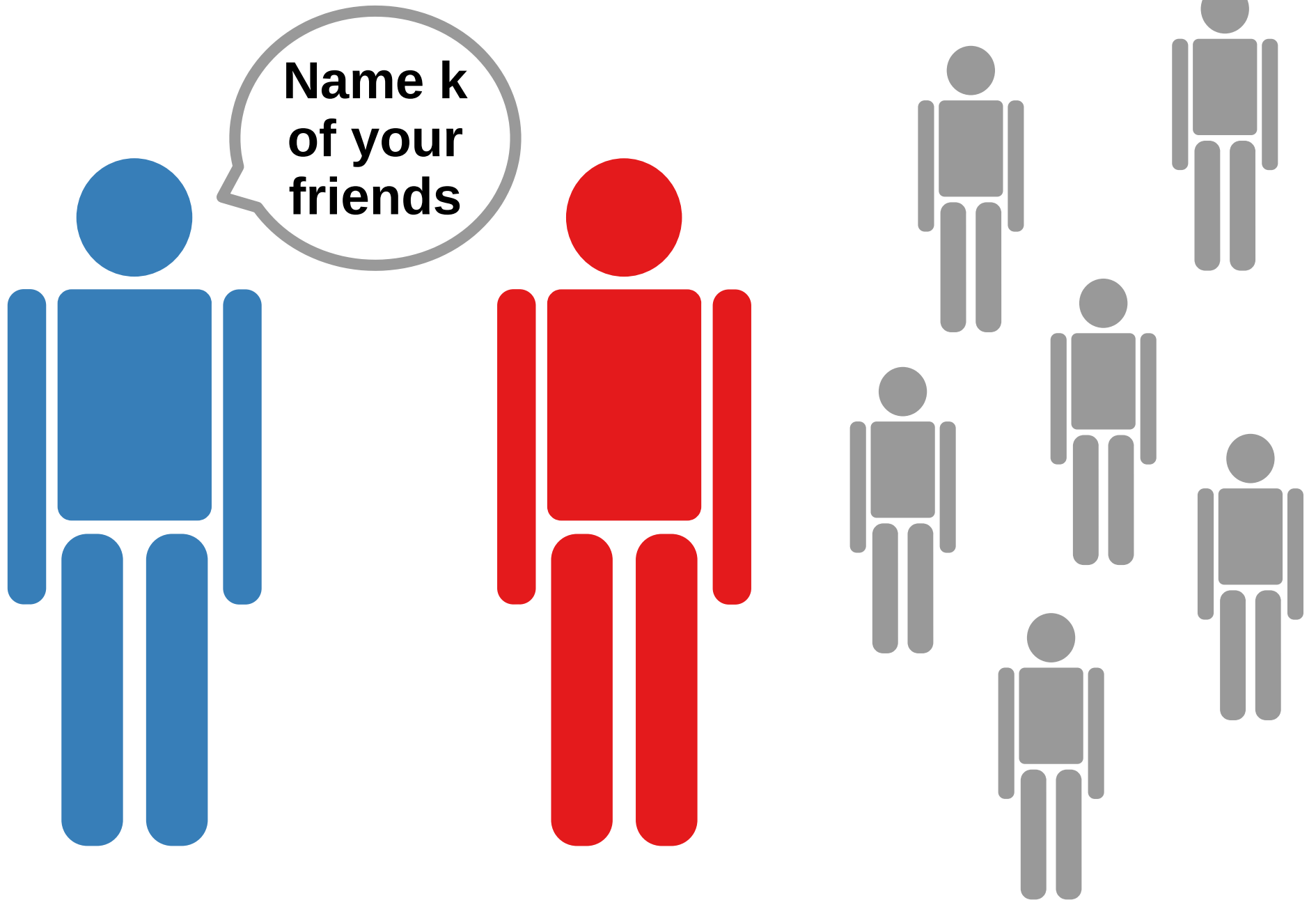
$$p_2 = \frac{20 * 1/2}{(50/1) + (20/2) + (10/3) + (8/4) + (7/5) + (5/6)}$$

$$p_2 = \frac{10}{67.5\bar{6}} \sim 0.148$$

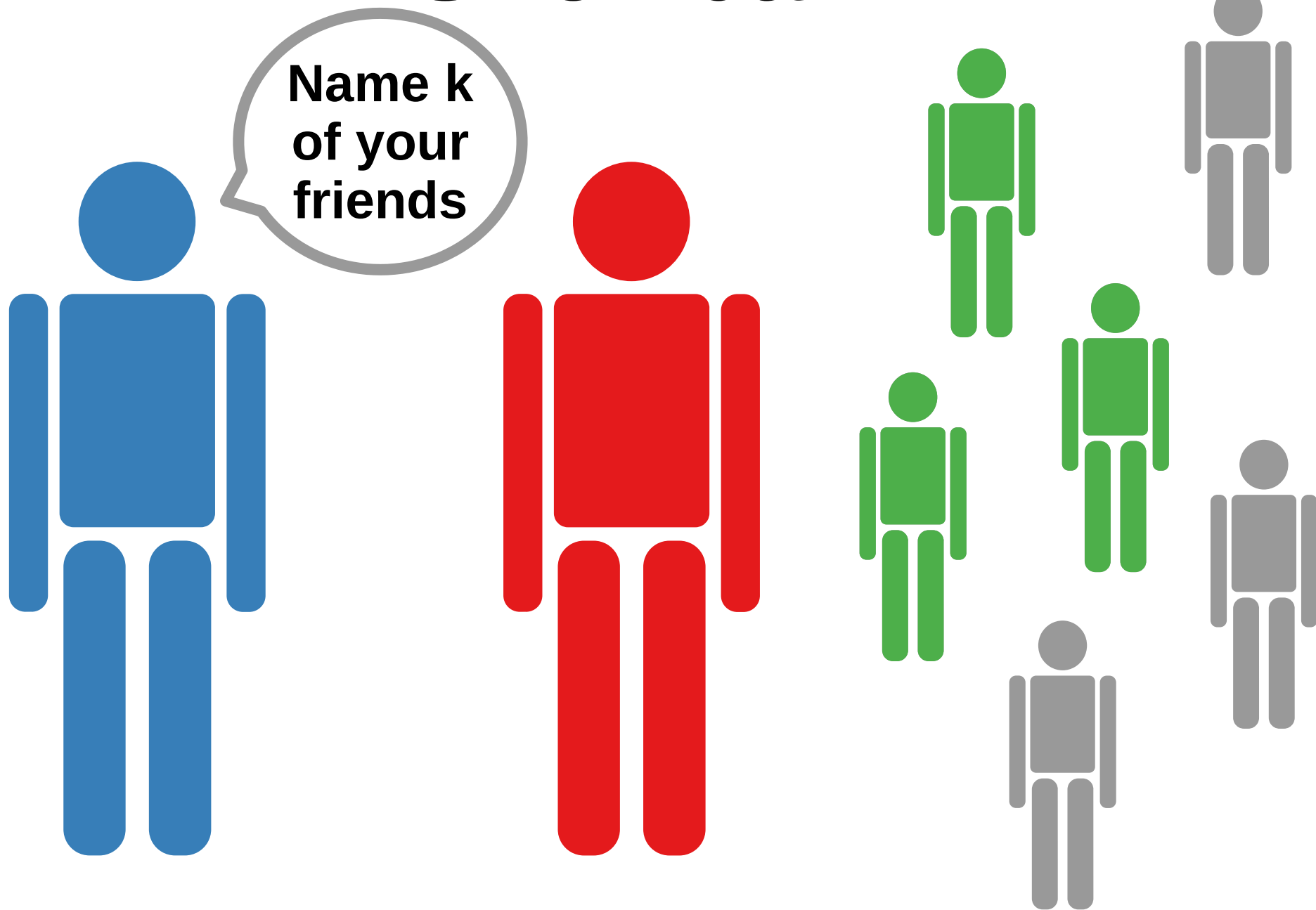
Snowball



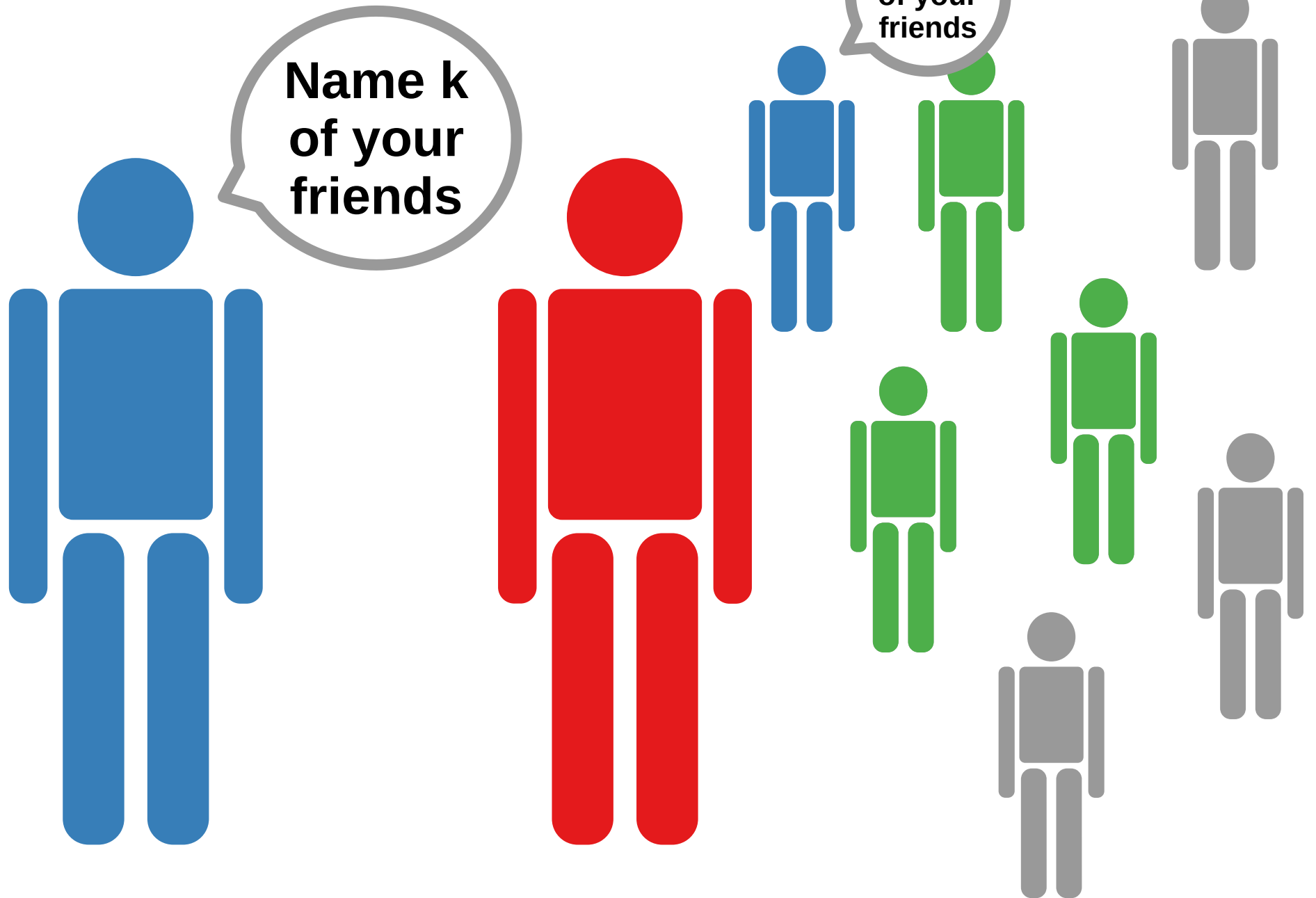
Snowball



Snowball



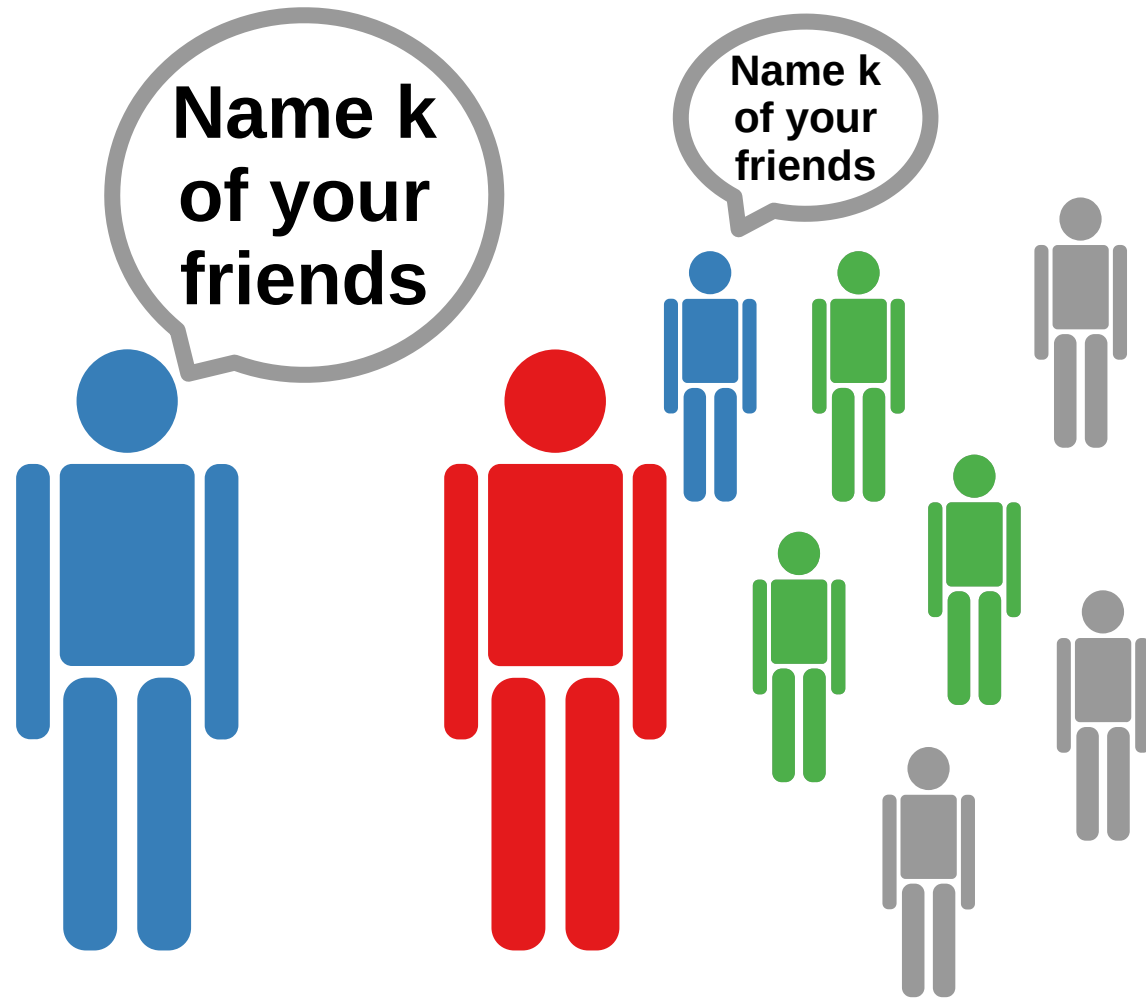
Snowball



Snowball: Advantages

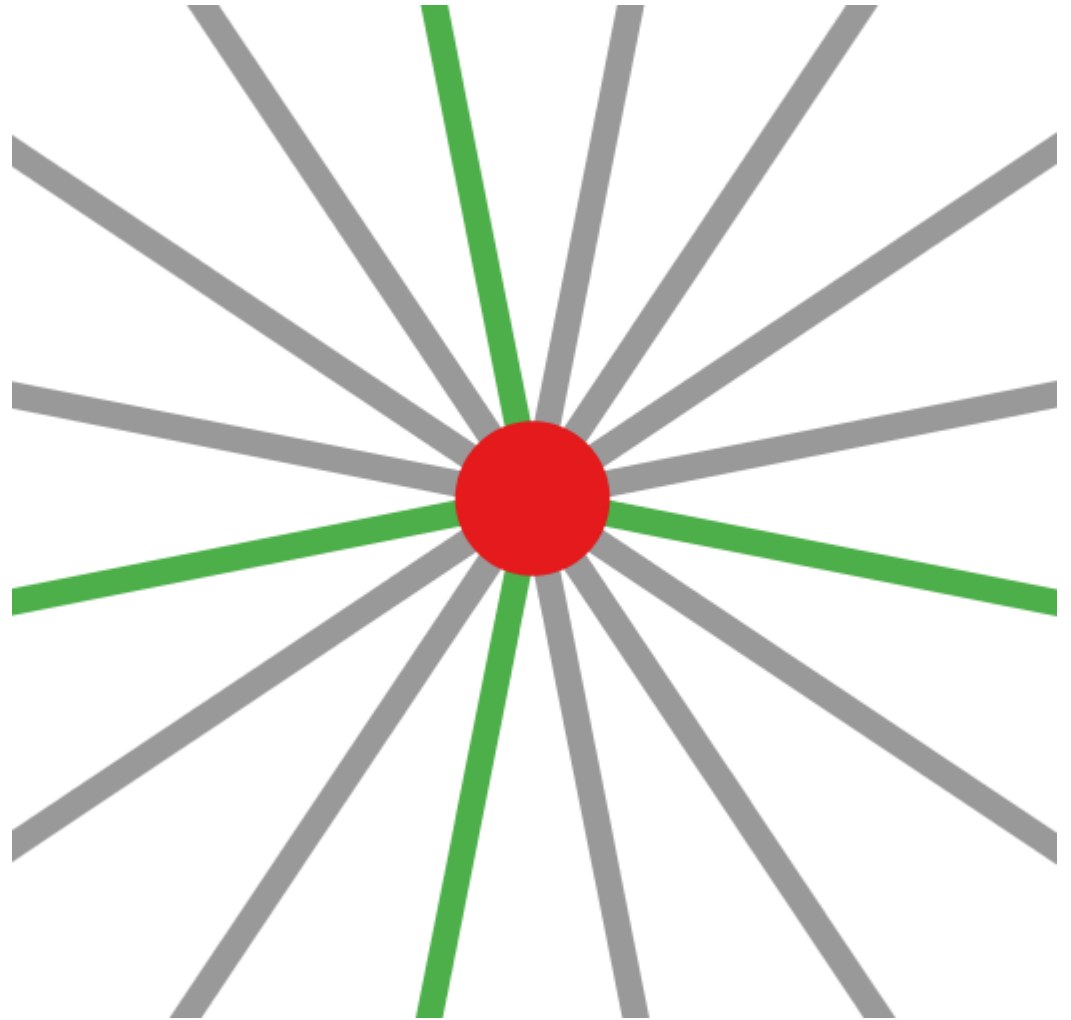
Snowball: Advantages

- Cheap in the physical world



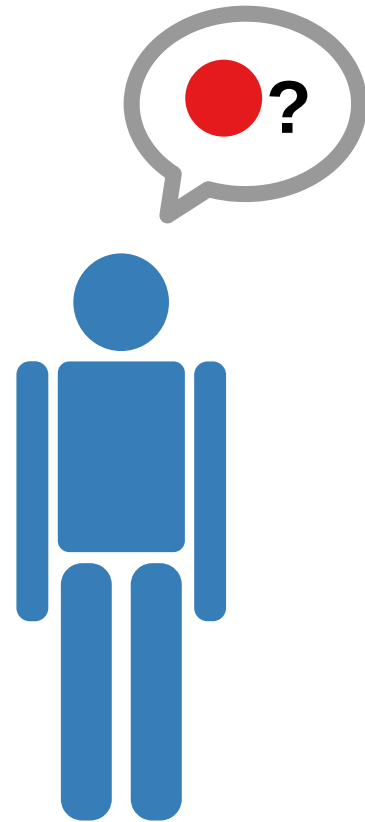
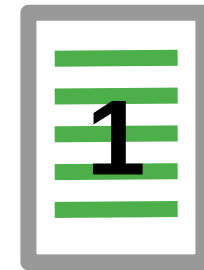
Snowball: Advantages

- Cheap in the physical world
- Smaller degree bias



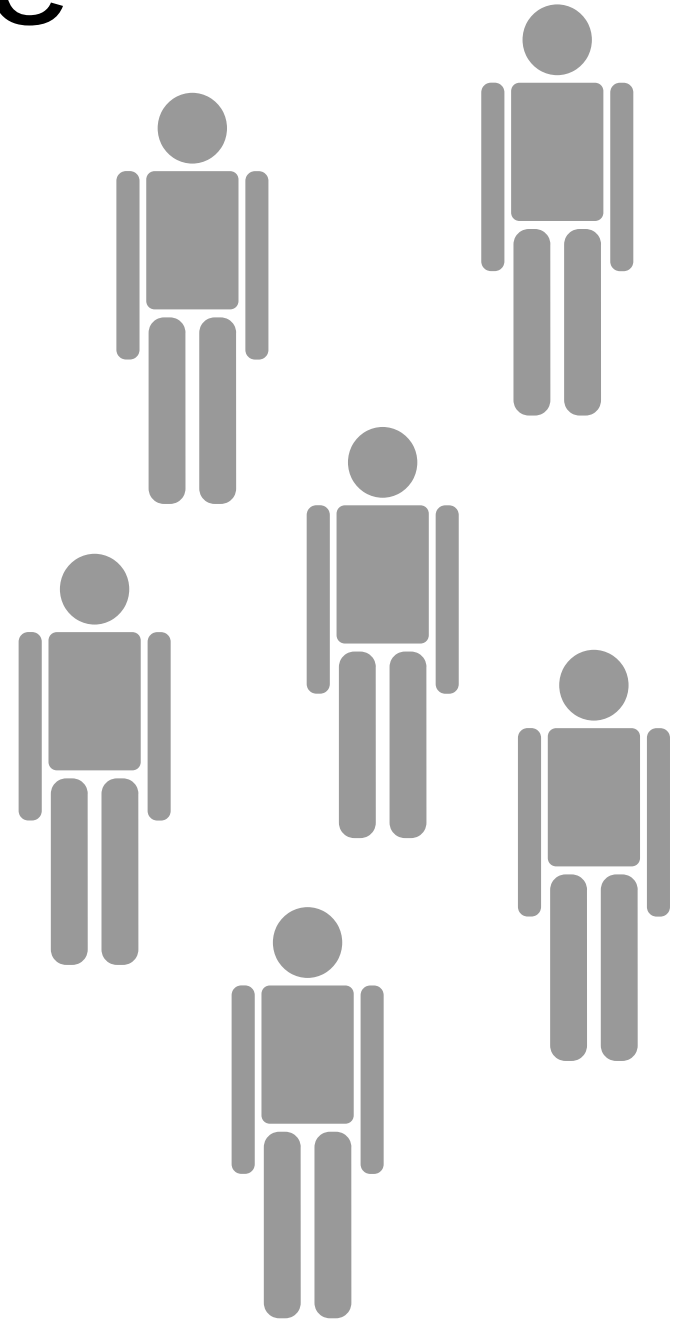
Snowball: Advantages

- Cheap in the physical world
- Smaller degree bias
- Works well with pagination



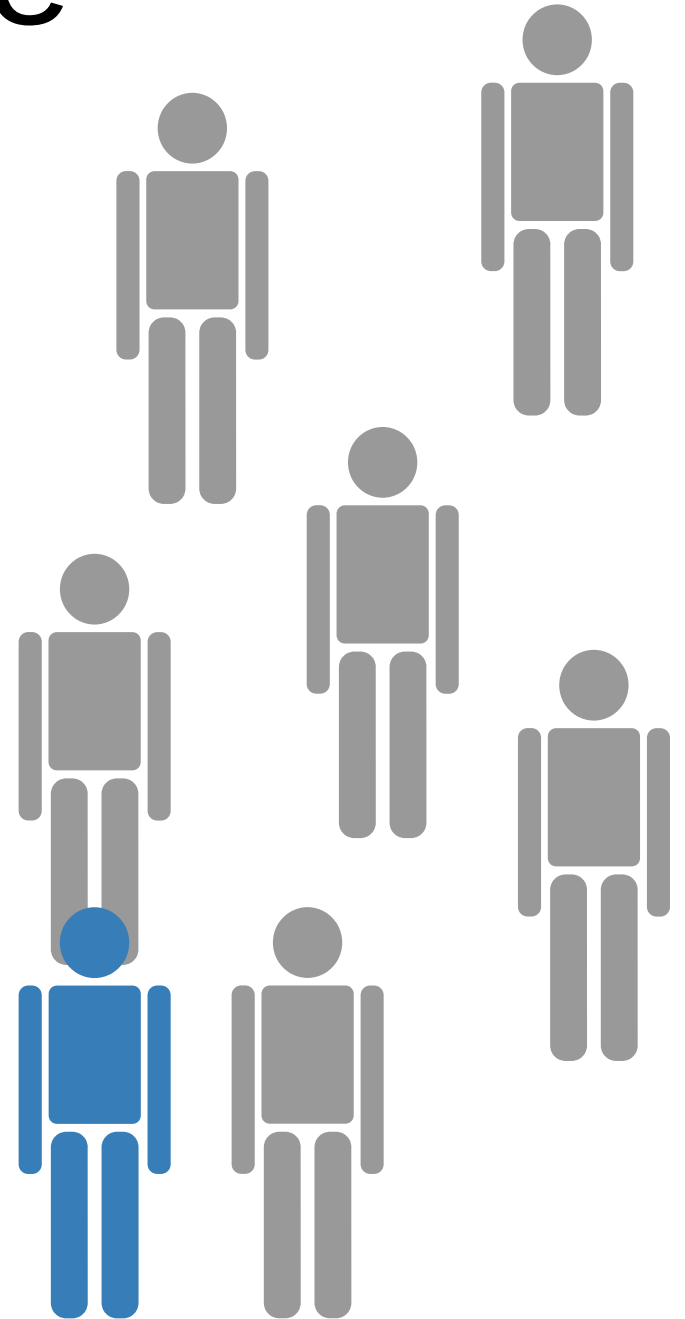
Forest Fire

**Name
all your
friends**



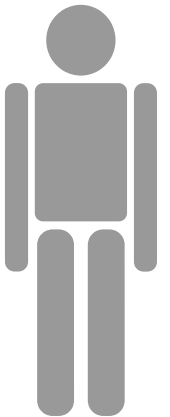
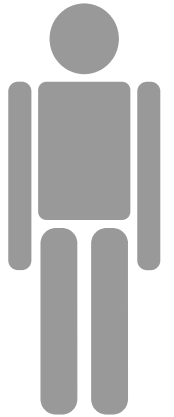
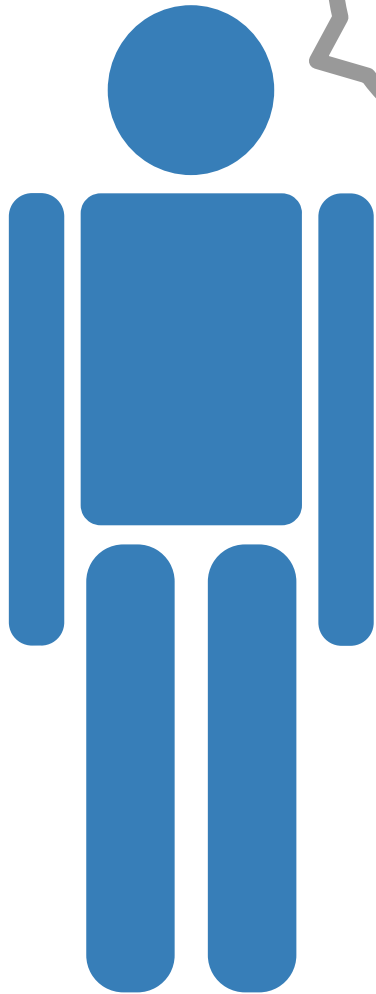
Forest Fire

**Name
all your
friends**



Forest Fire

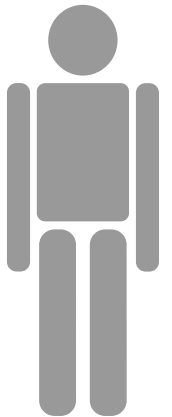
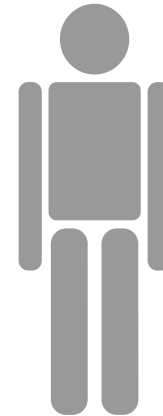
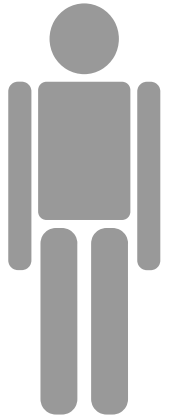
**Name
all your
friends**



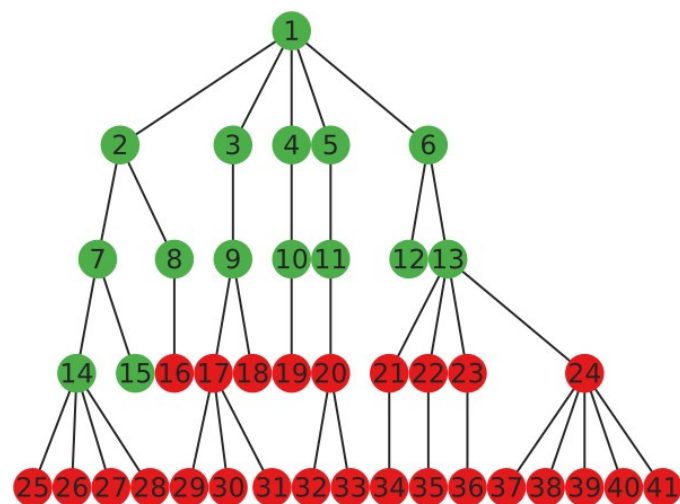
**Name
all your
friends**

**Name
all your
friends**

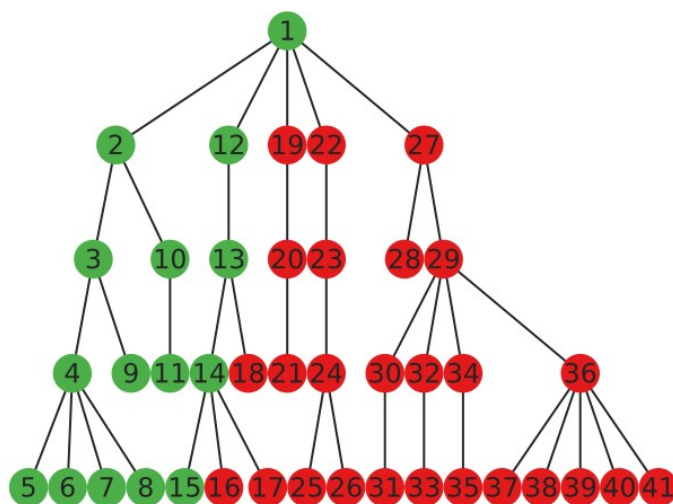
**Name
all your
friends**



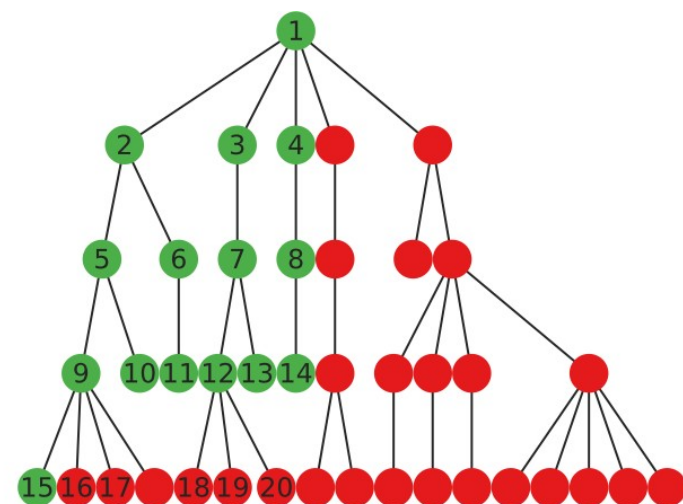
The Network Sampling Zoo



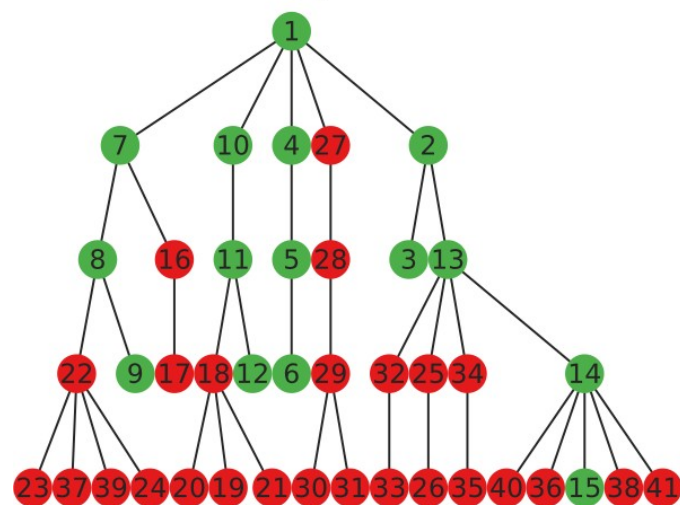
(a) BFS



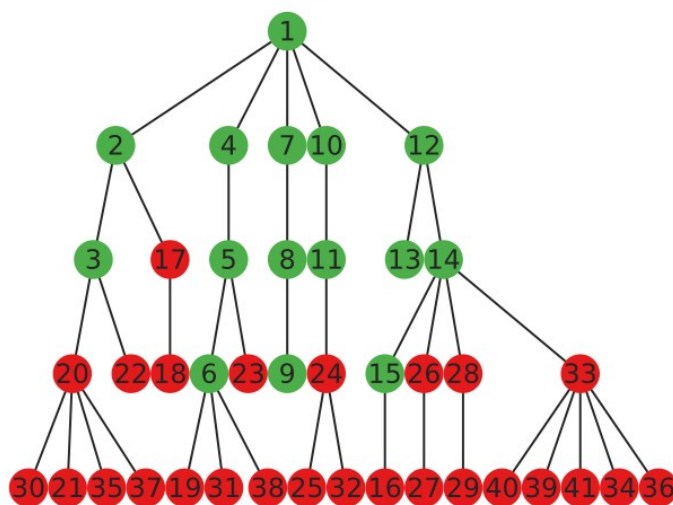
(b) DFS



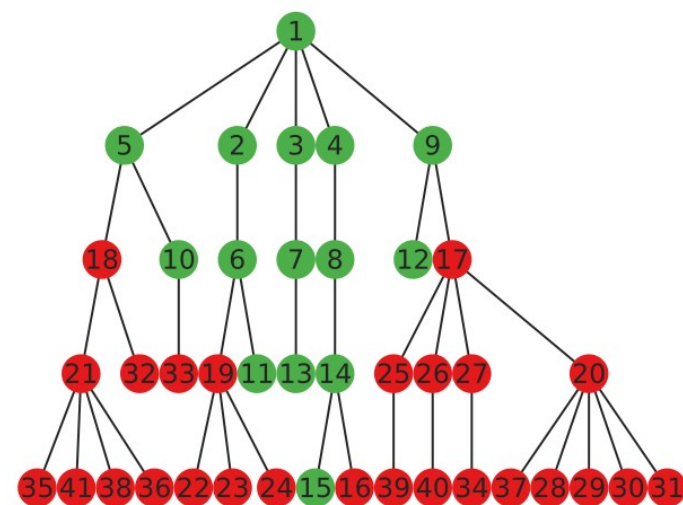
(c) Snowball



(d) Random Walk



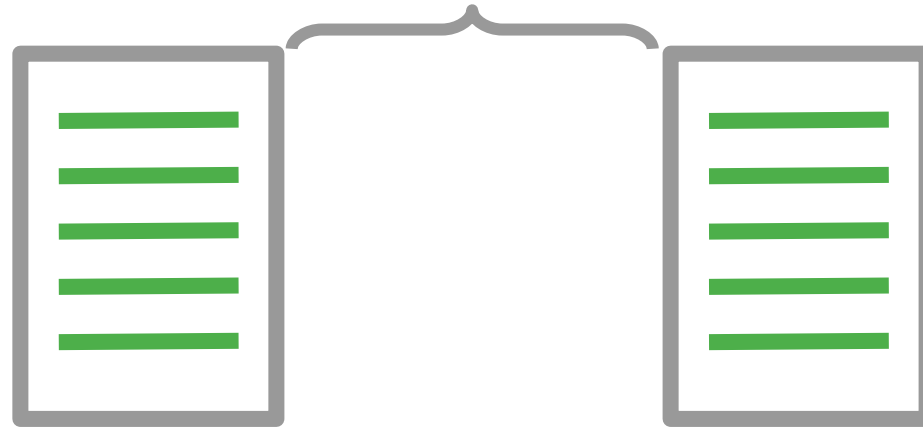
(e) MHRW



(f) Forest Fire

(ii) Are there understudied real world obstacles that should make us reconsider how we choose the best sampling strategy?

Social Media APIs

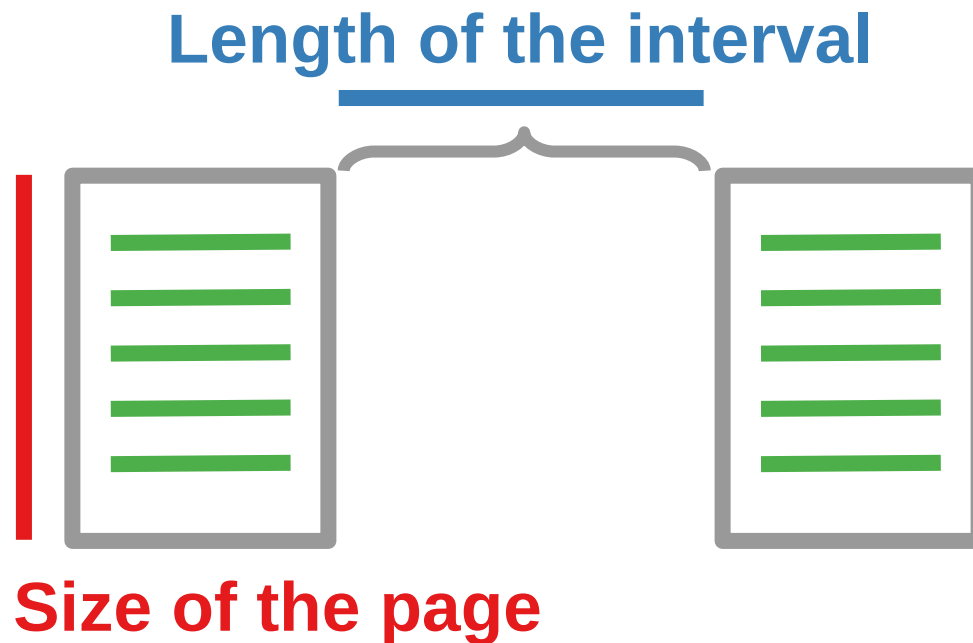


Social Media APIs

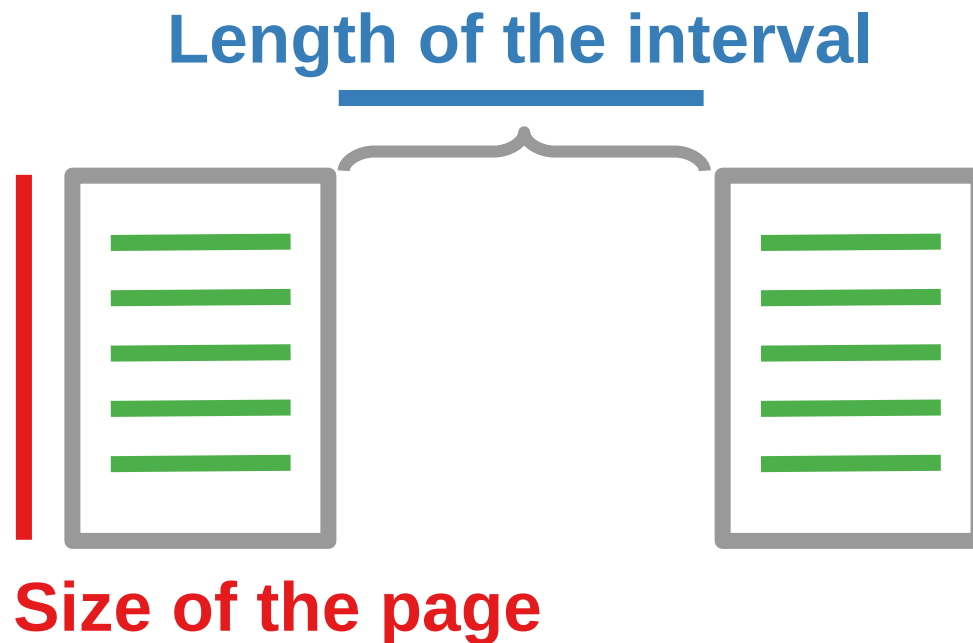


Size of the page

Social Media APIs



Social Media APIs



(...latency)

Pagination Paradox

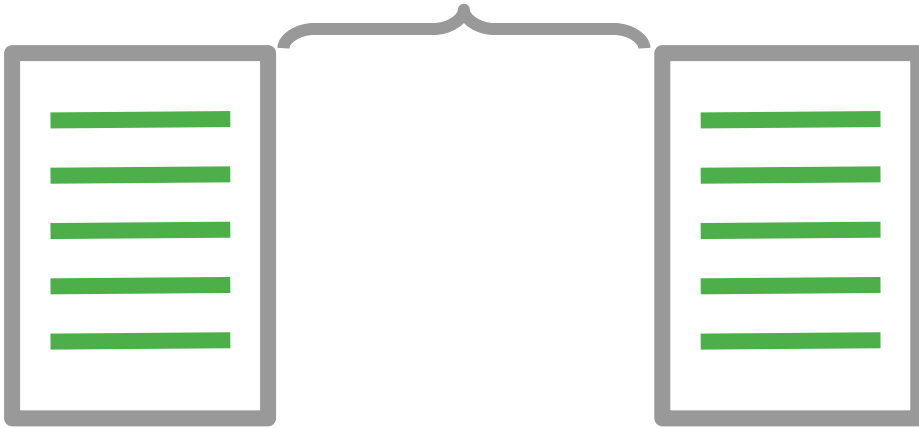


Pagination Paradox



- Edges per page:
100
- Seconds between
queries: 2
- 50 edges / sec

Pagination Paradox

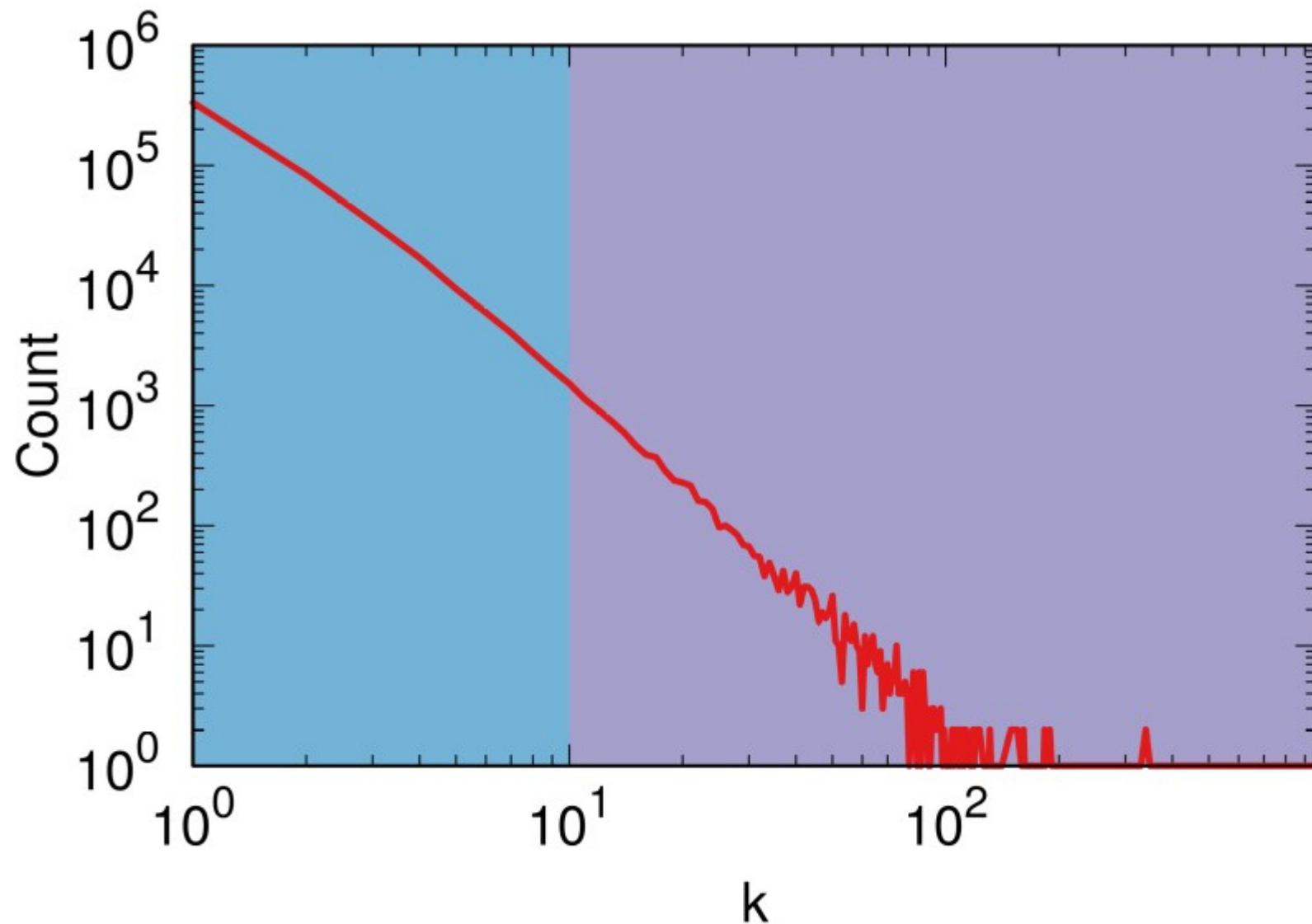


- Edges per page: 100
- Seconds between queries: 2
- 50 edges / sec

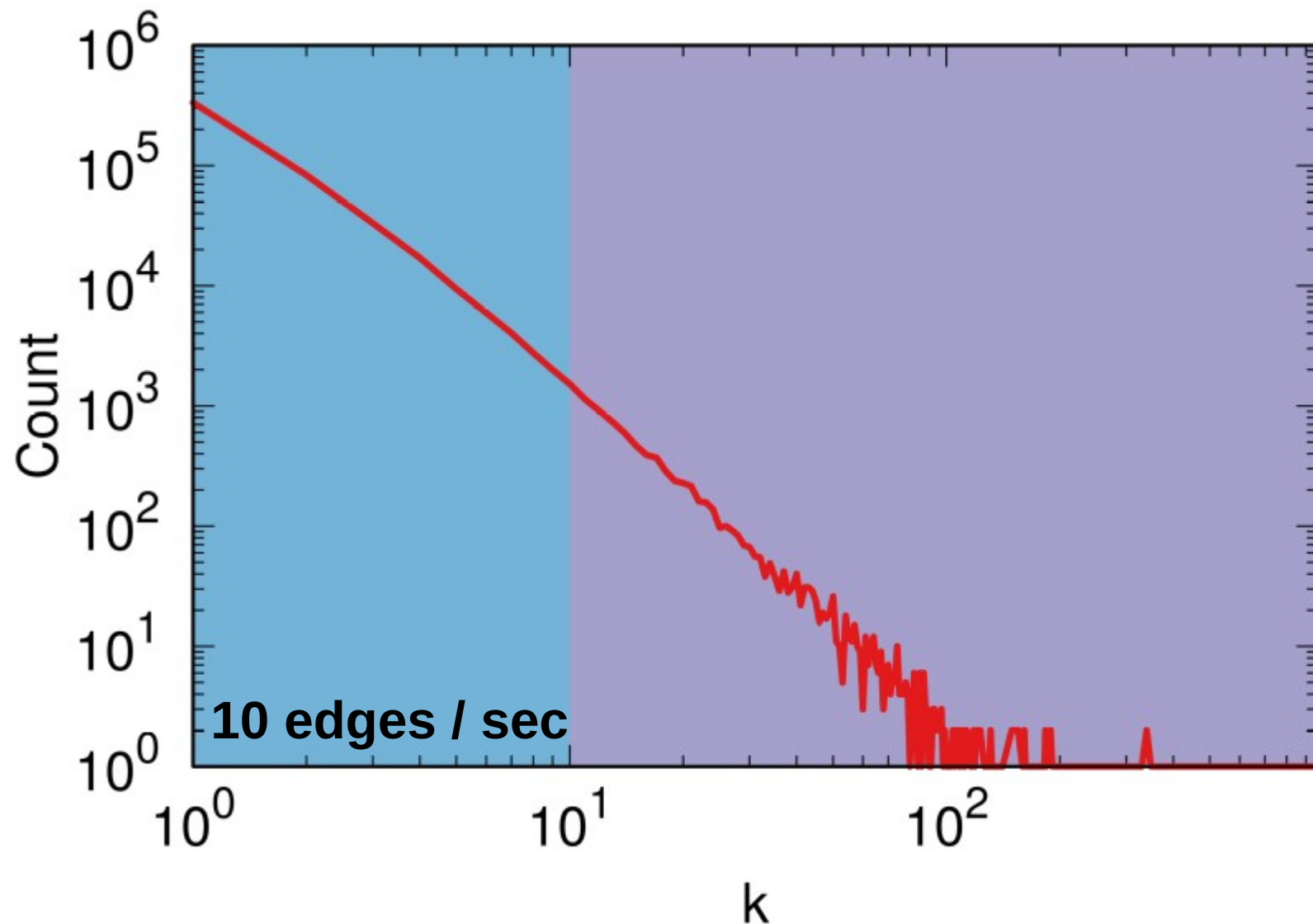


- Edges per page: 10
- Seconds between queries: 1
- 10 edges / sec

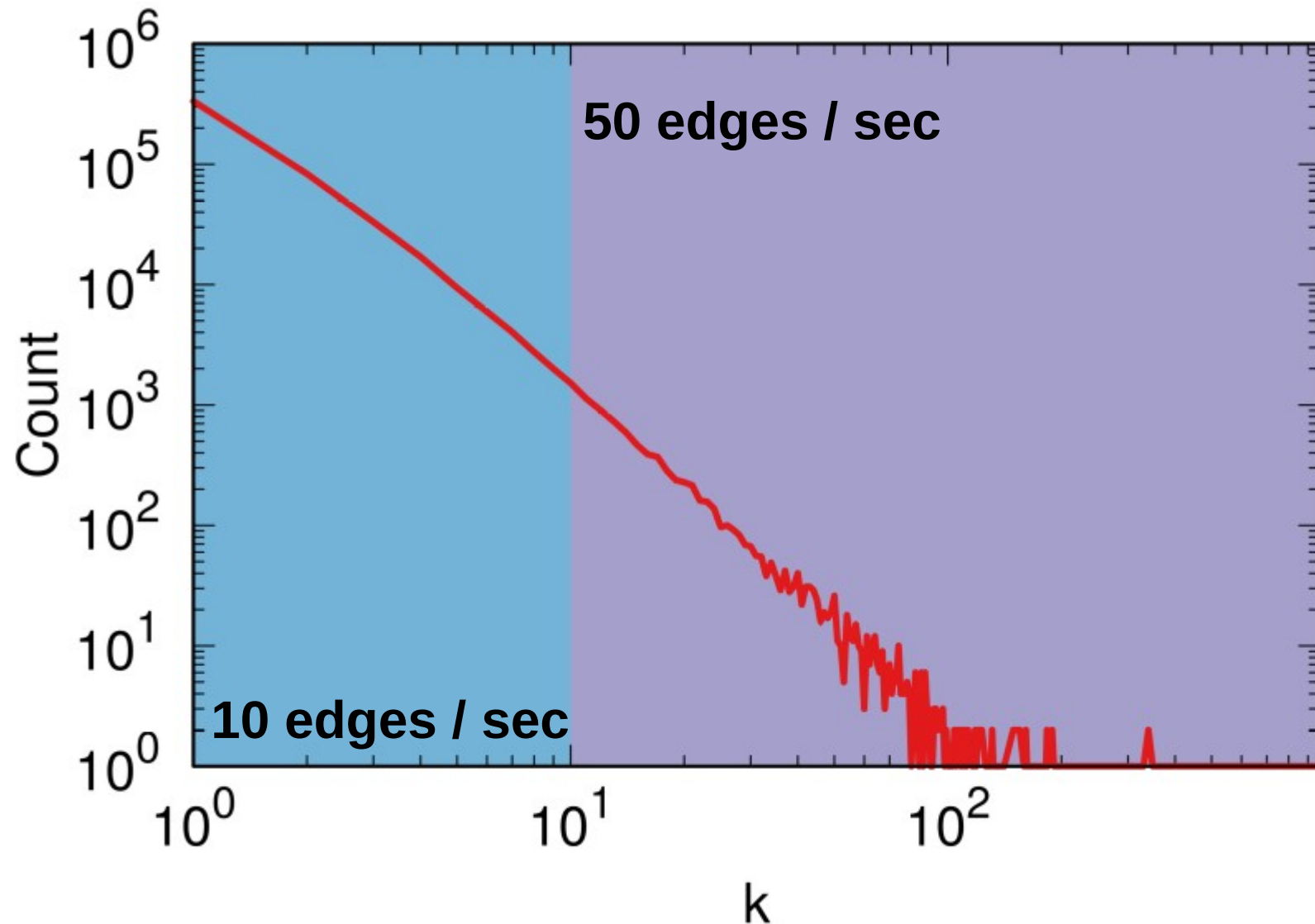
Pagination Paradox



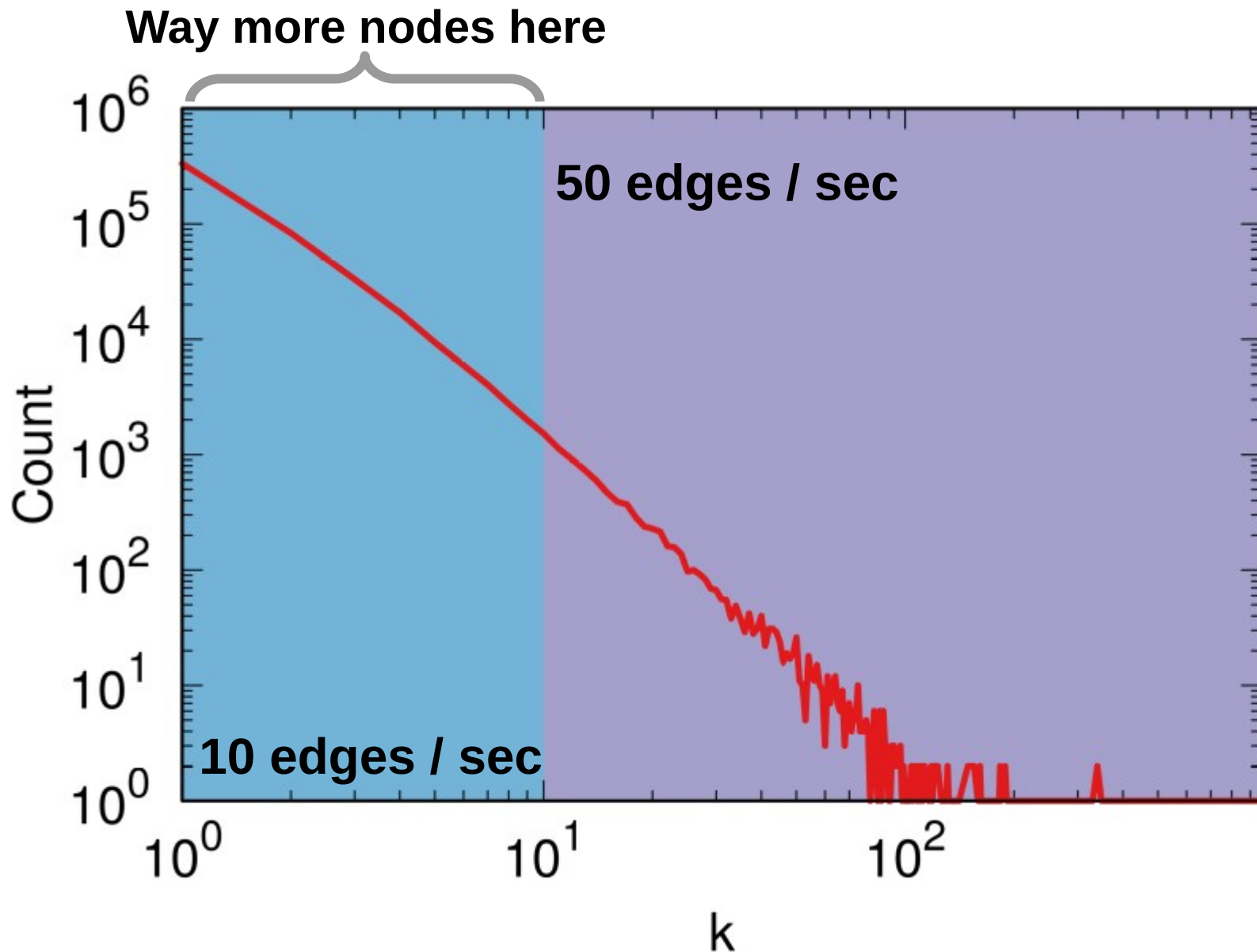
Pagination Paradox



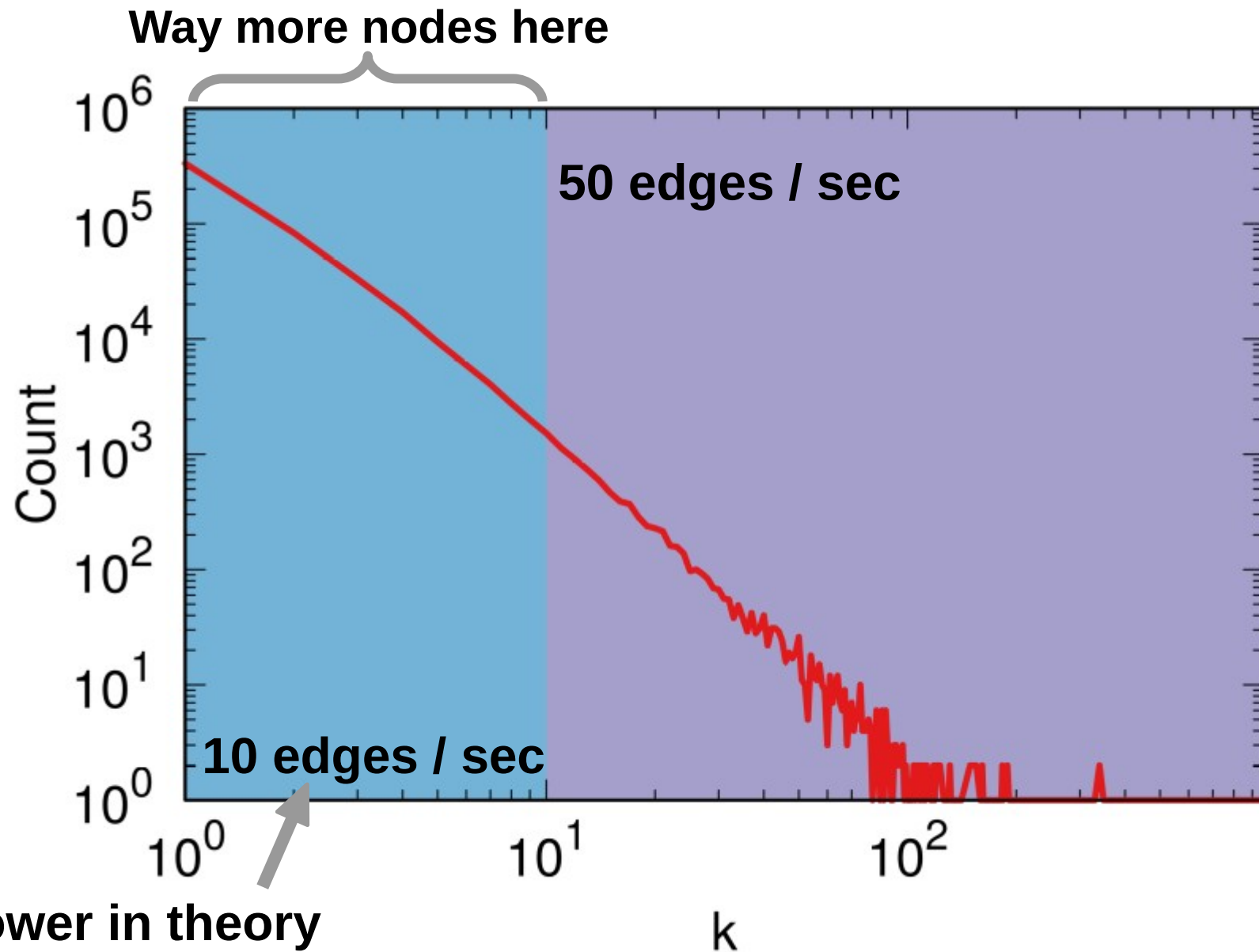
Pagination Paradox



Pagination Paradox



Pagination Paradox

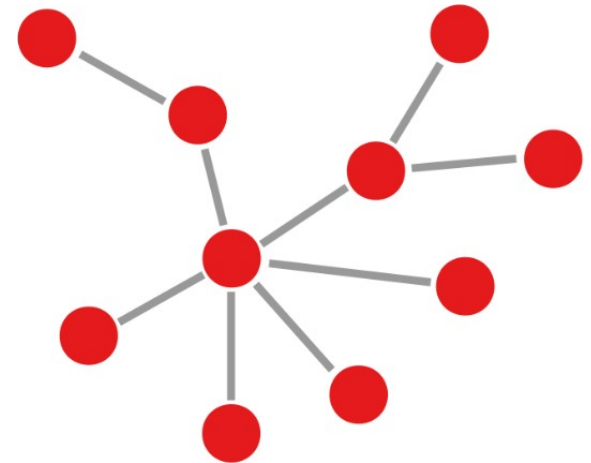


Benchmark Setup

- Three types of topologies:
 - Barabasi-Albert
 - Small World
 - LFR Benchmark

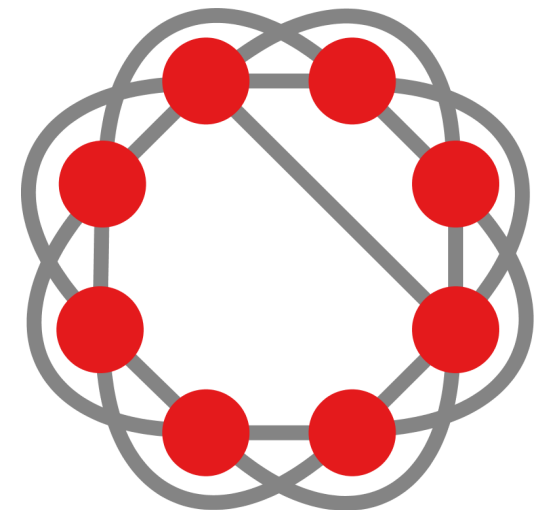
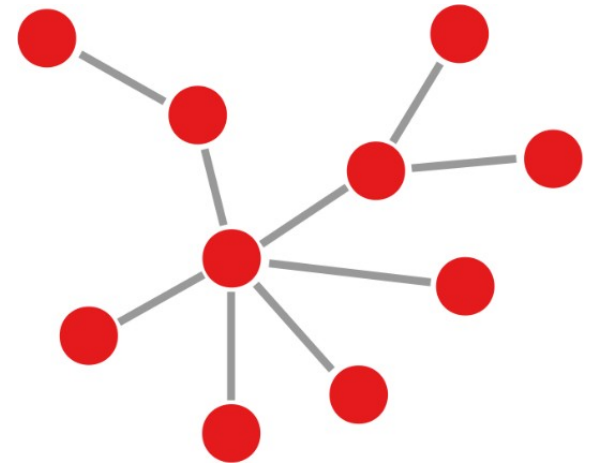
Benchmark Setup

- Three types of topologies:
 - Barabasi-Albert
 - Small World
 - LFR Benchmark



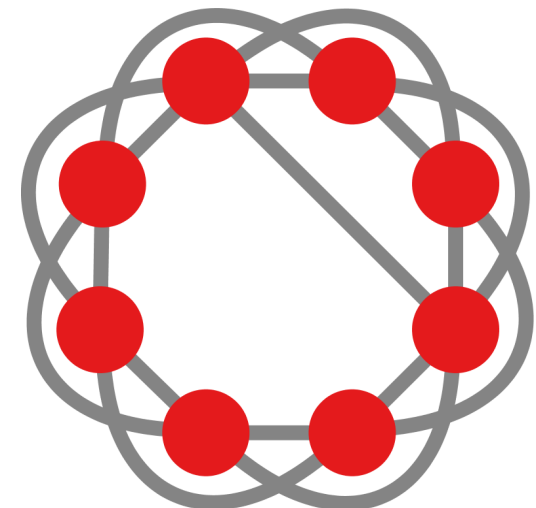
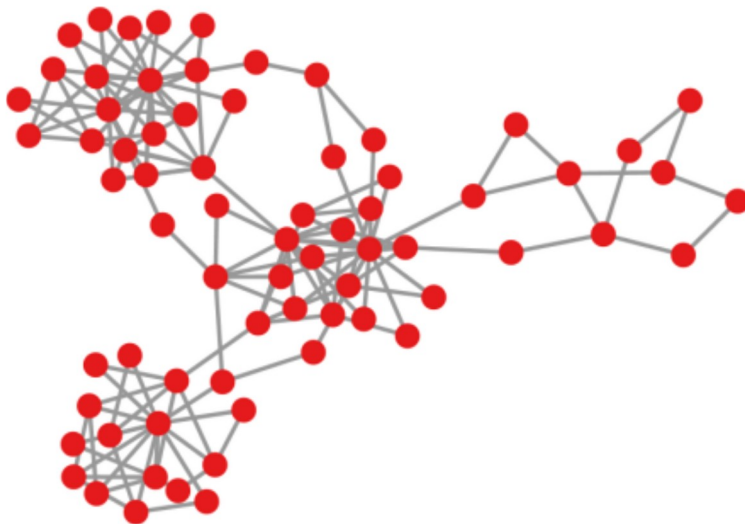
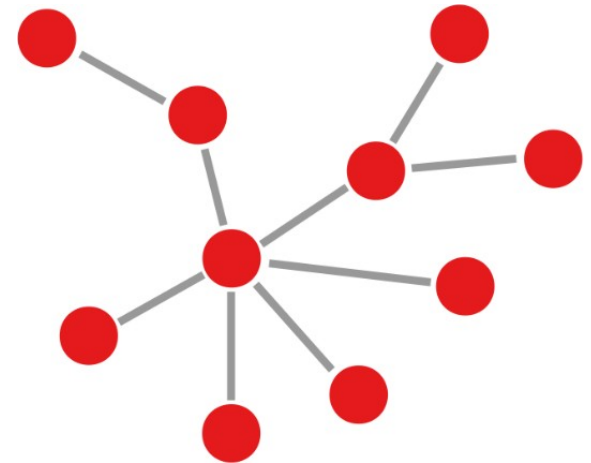
Benchmark Setup

- Three types of topologies:
 - Barabasi-Albert
 - Small World
 - LFR Benchmark



Benchmark Setup

- Three types of topologies:
 - Barabasi-Albert
 - Small World
 - LFR Benchmark



Benchmark Setup

- Six API systems from real social media:
 - Flickr
 - Lastfm
 - Twitter
 - Youtube
 - Tumblr
 - Google+

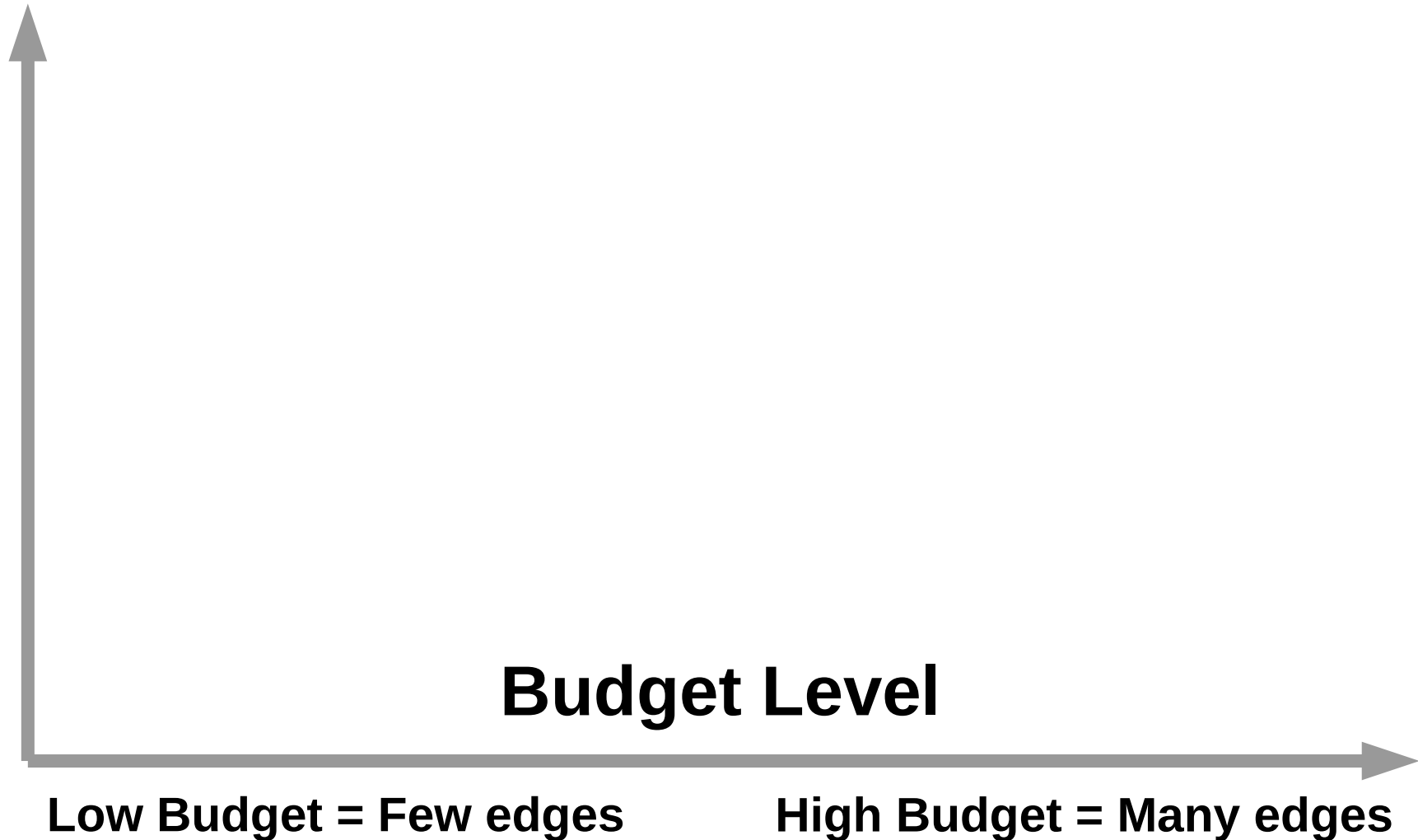
Benchmark Setup

- Different objectives:
 - Degree Distribution
 - Assortativity / Disassortativity
 - Centrality
 - Reciprocity

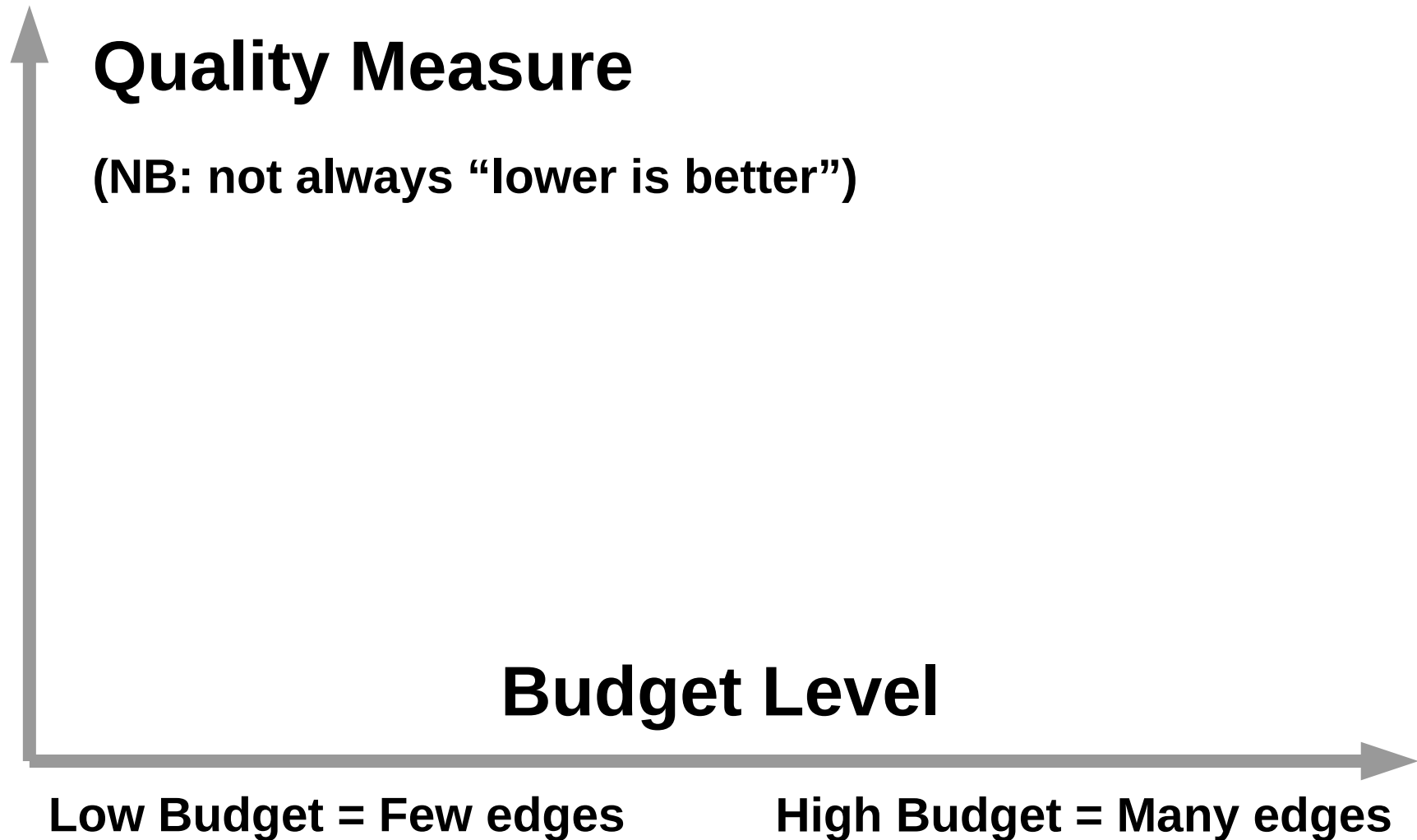
Benchmark Setup



Benchmark Setup



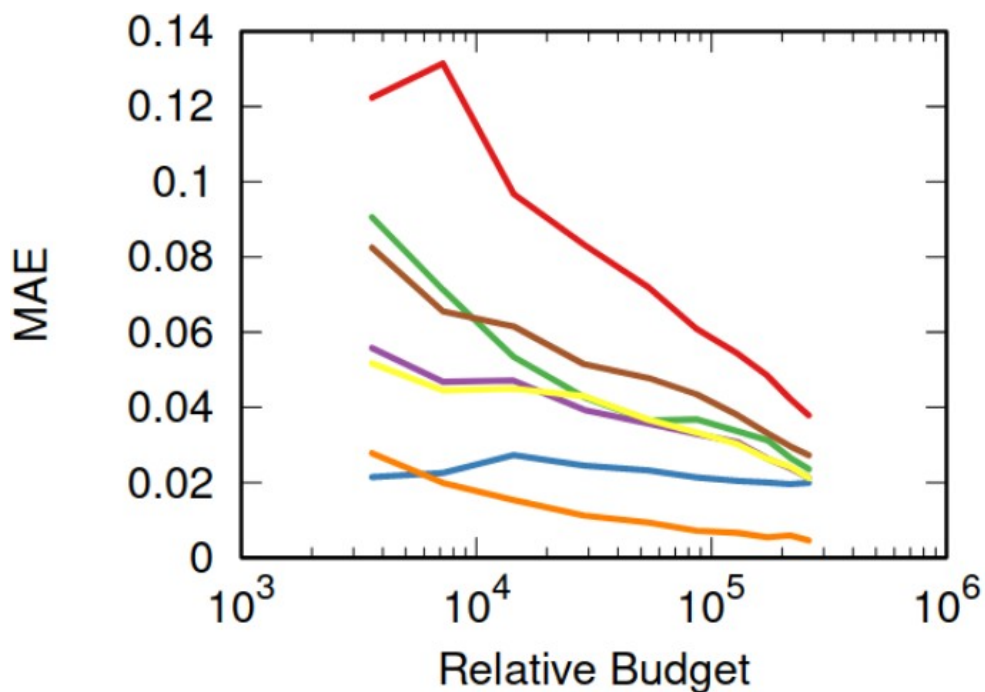
Benchmark Setup



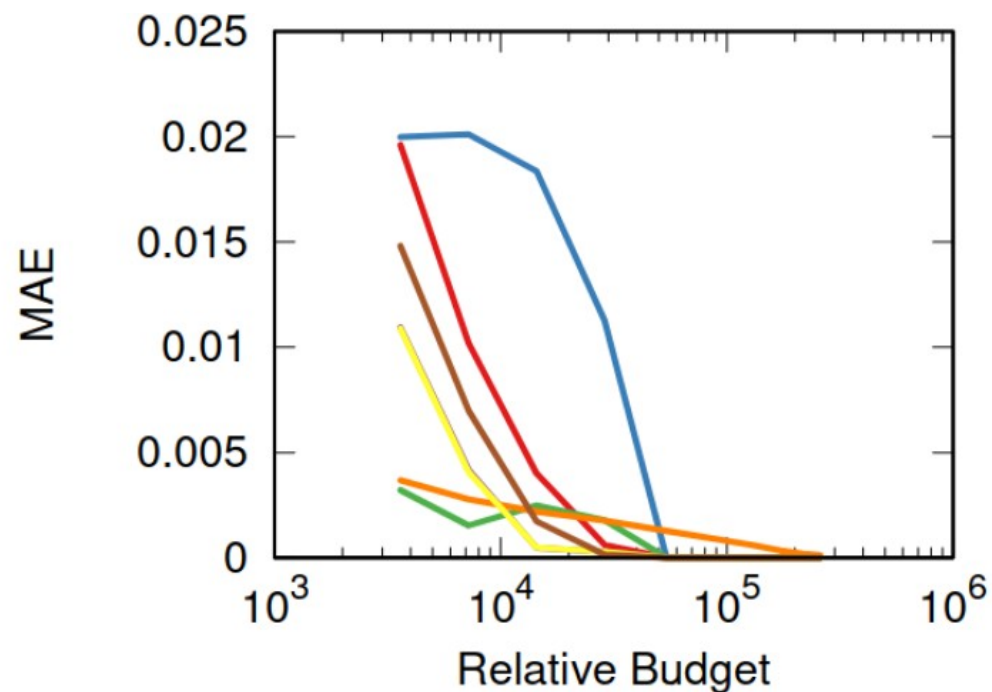
Disassortativity MAE

(lower is better)

Tumblr



LastFM

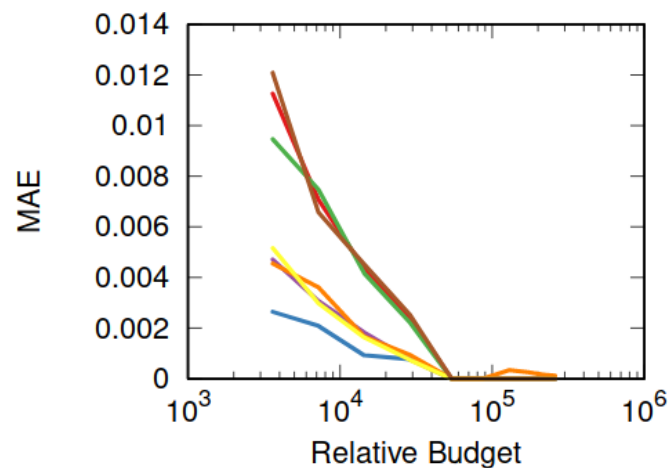


BFS — DFS — SBS — RW — MHRW — RWRW — FF —

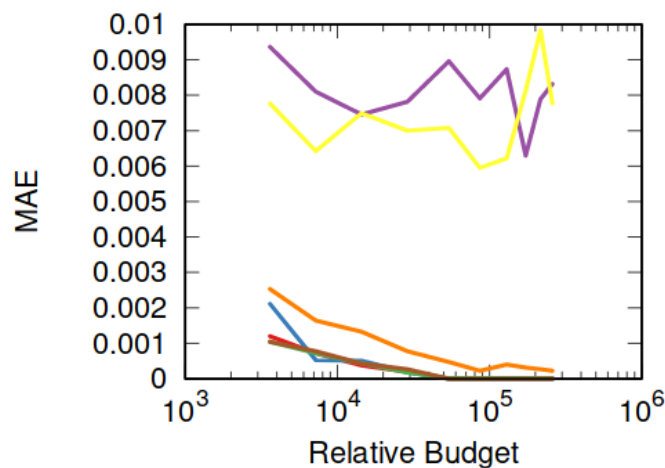
Assortativity MAE

(lower is better)

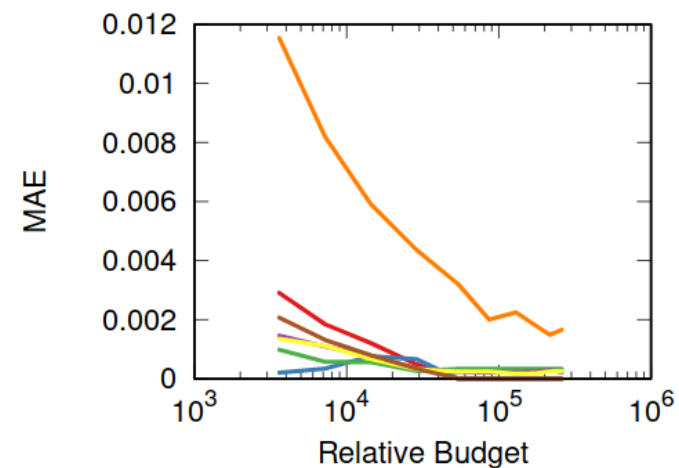
SW



B-A



LFR

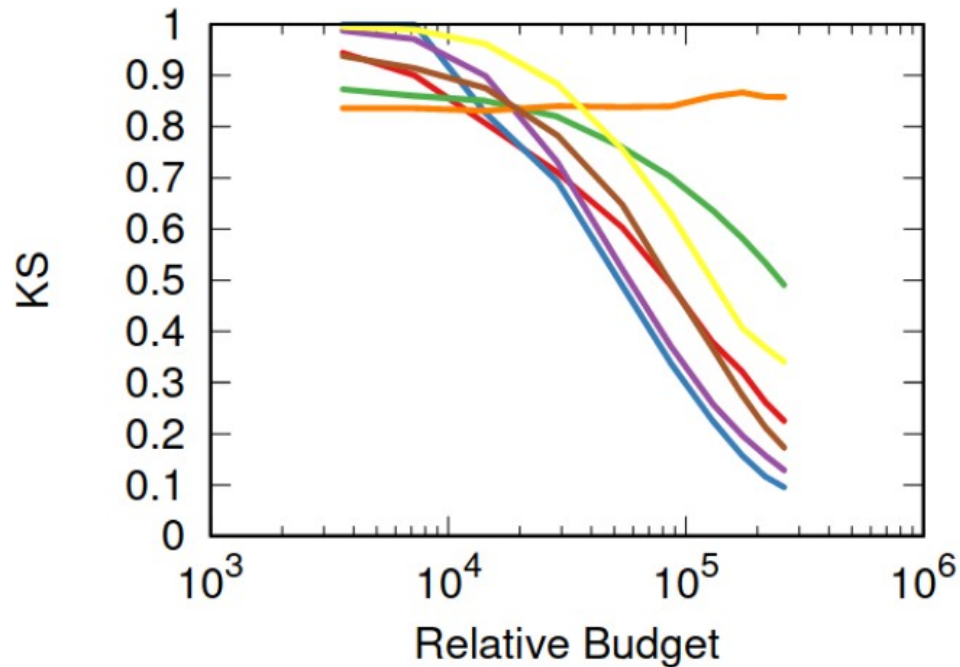


BFS DFS SBS RW MHRW RWRW FF

Budget Levels

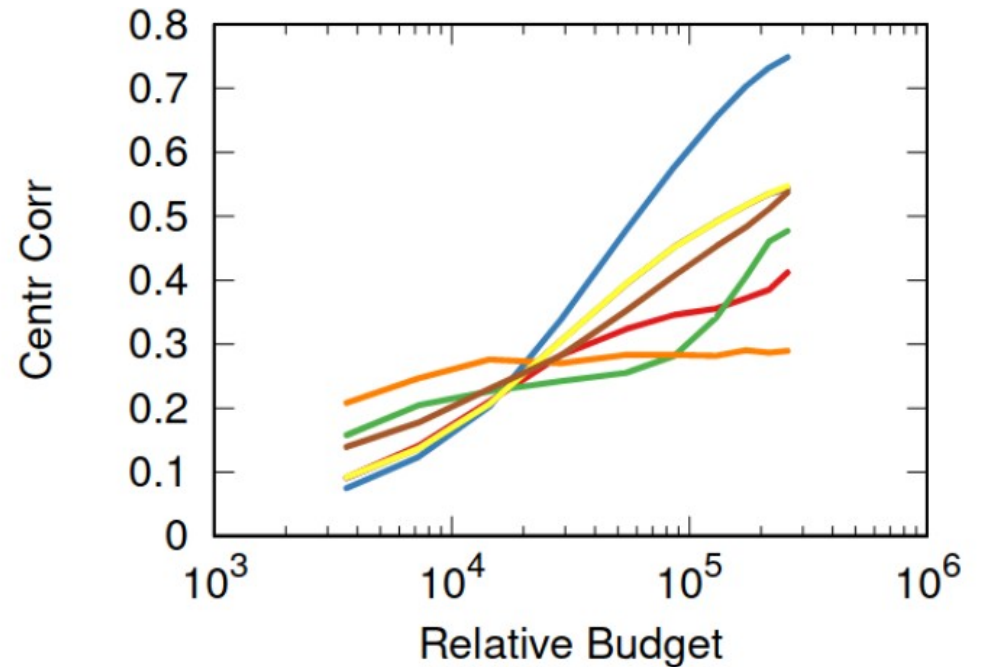
Degree Distribution

(lower is better)



Centrality Correlation

(higher is better)



BFS — DFS — SBS — RW — MHRW — RWRW — FF —

Conclusion

Conclusion

- We have to sample

Conclusion

- We have to sample
- We have good theory...

Conclusion

- We have to sample
- We have good theory...
- ...for the case of infinite time and paging sizes

Conclusion

- We have to sample
- We have good theory...
- ...for the case of infinite time and paging sizes
- Which is not realistic

Conclusion

- We have to sample
- We have good theory...
- ...for the case of infinite time and paging sizes
- Which is not realistic
- Realistic constraints paint a critical picture

Thanks

Benchmarking API Costs of Network Sampling Strategies

Michele Coscia & Luca Rossi

mcos@itu.dk lucr@itu.dk

<http://www.michelecoscia.com>