Finding the influential users in the Twitter network

HUSSAIN JAGIRDAR (17CS60R83)
NIKHIL AGARWAL (17CS60R70)
SHAH SMIT KETANKUMAR (17CS60R72)

MENTOR
AYAN KUMAR BHOWMICK
DR BIVAS MITRA

Content

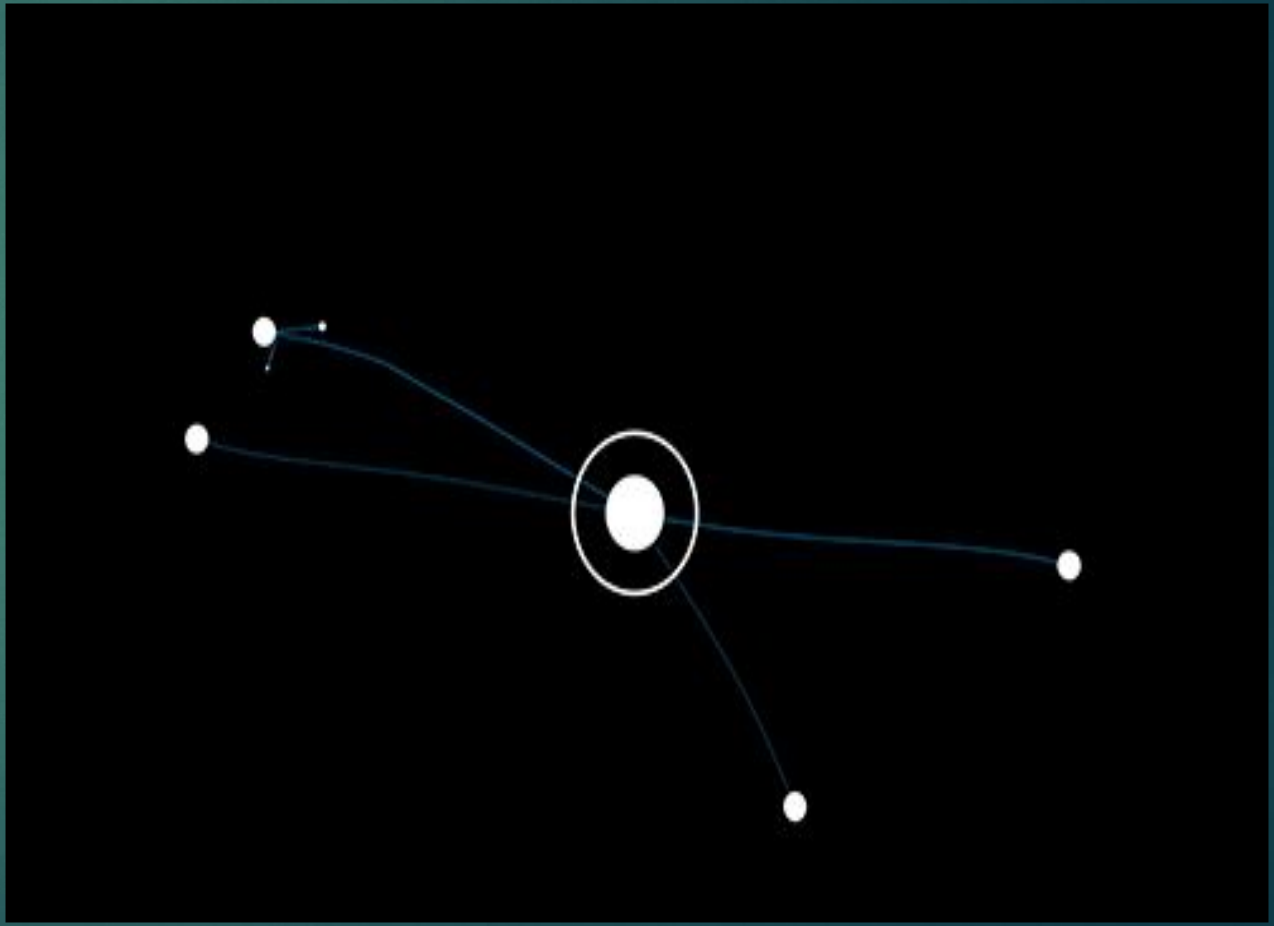
- Problem Description
- Objective
- Progress Made
  - Data Collection
  - Methods Implemented
  - Evaluation Metrics
- Plan of action

Problem Description

Identifying Influential users

Need – identifying Influential users

- The viral marketing to maximize ROI (Return of Investment).
- Targeting the influential nodes to transfer information during epidemics and natural calamities.
- Search expertise/tweets recommendation
- Trust/information propagation.



Influence in Social Media

- Online communication is the new way to receive information
- Influence : the power or capacity of causing an effect in indirect or intangible ways.
- Twitter:

  - Can share messages of length up to **280 characters**

  - People can retweet too (a reposted or forwarded message)

  - Causes information diffusion over the global follower network

  - The final reach may depend on tweets posted by certain influential users

Objective

- Finding the top influential users in a twitter network based on the static as well as temporal methodologies.

- Comparative study of performance between these methods.

Data Collected

- Two publicly available tweet datasets
- Algeria and Egypt datasets connected to the Arab-Spring Movement

  - collection of tweets (tweet-ids) and users who posted them.

Dataset	#Tweets	#Retweets	#Cascades	#ActiveUsers	Maximum size of cascade
Algeria	65268	17269	5730	8814	980
Egypt	671417	188090	67539	13882	432

Dataset

```
tweet-id : user1 user2 user3 .. usern
tweet-id : time1 time2 time3 .. timen

ABC : Smit  Hussain Nikhil
ABC : 10    21      41
ABC : 10    11      20      (Time interval)
```

Methods Implemented

Methods used to find static influential nodes

Degree Centrality

Page-rank Centrality

MCDWE score

Borda Count

K-truss decomposition

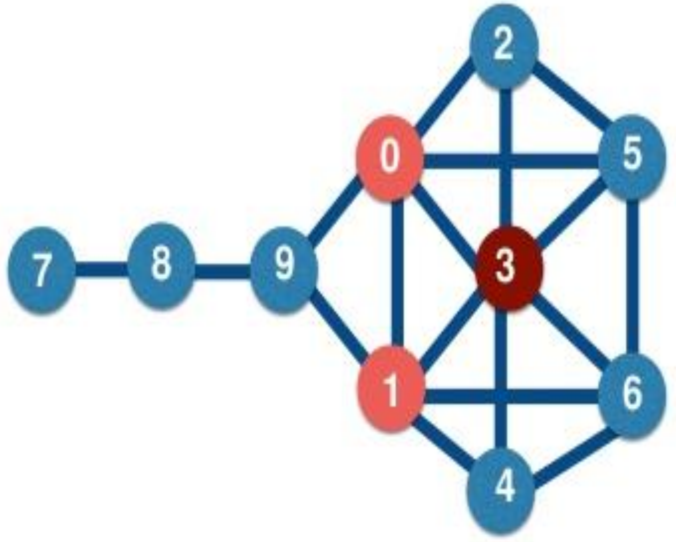
Methods Implemented

Degree Centrality

- Degree is a simple centrality measure that counts how many neighbors a node has.
- If the network is directed, we have two versions of the measure: in-degree is the number of in-coming links, or the number of predecessor nodes; out-degree is the number of out-going links, or the number of successor nodes.

Finding score for influential users through degree centrality

3	0.666
0	0.555
1	0.555
5	0.444
6	0.444
2	0.333
4	0.333
9	0.333
8	0.222
7	0.111



nx.degree_centrality(G)

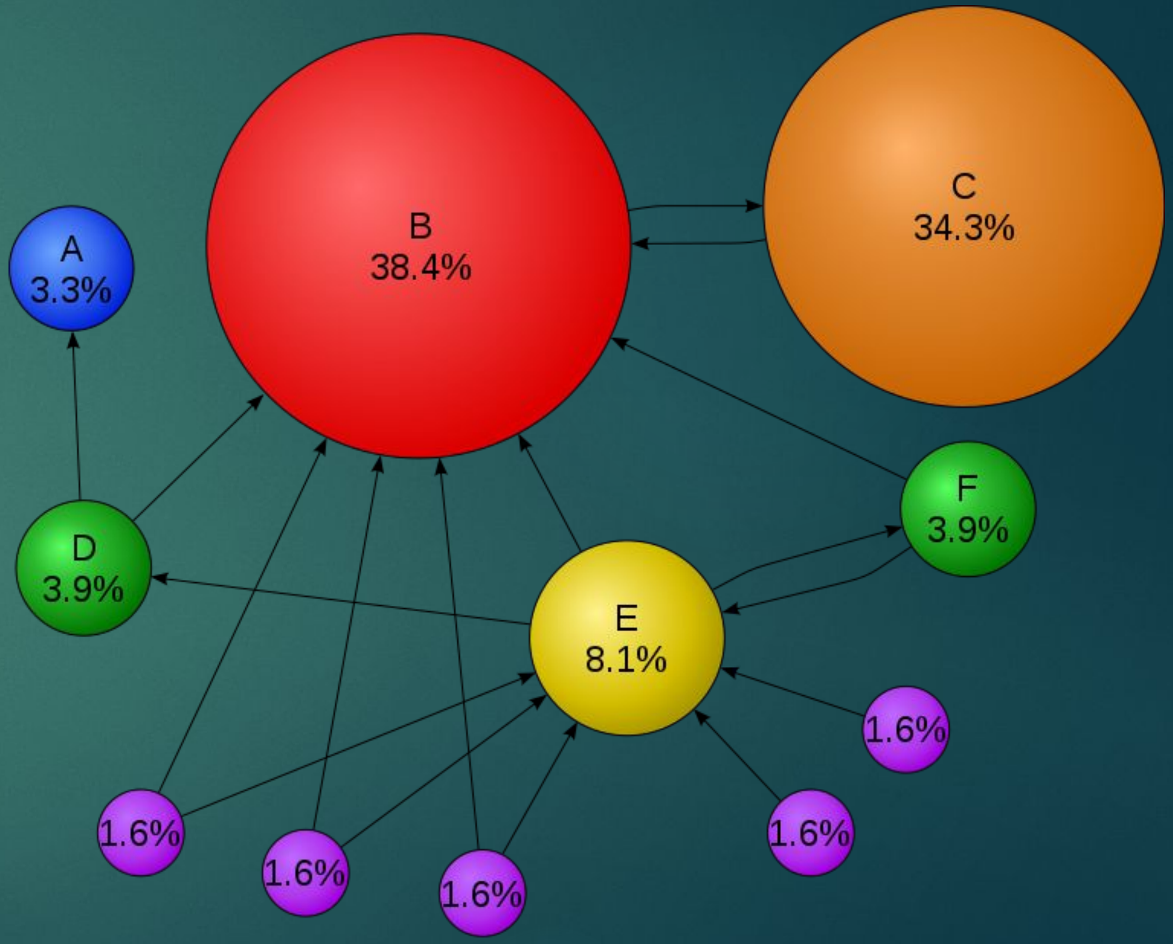
= number of edges directly connected to n

Methods Implemented

Page-rank Centrality

- PageRank works by counting the number and quality of connection to a user to determine a rough estimate of how important/influential the user is.

- The underlying assumption is that more important/influential users are likely to receive more links from other users.



Methods Implemented

MCDWE ranking

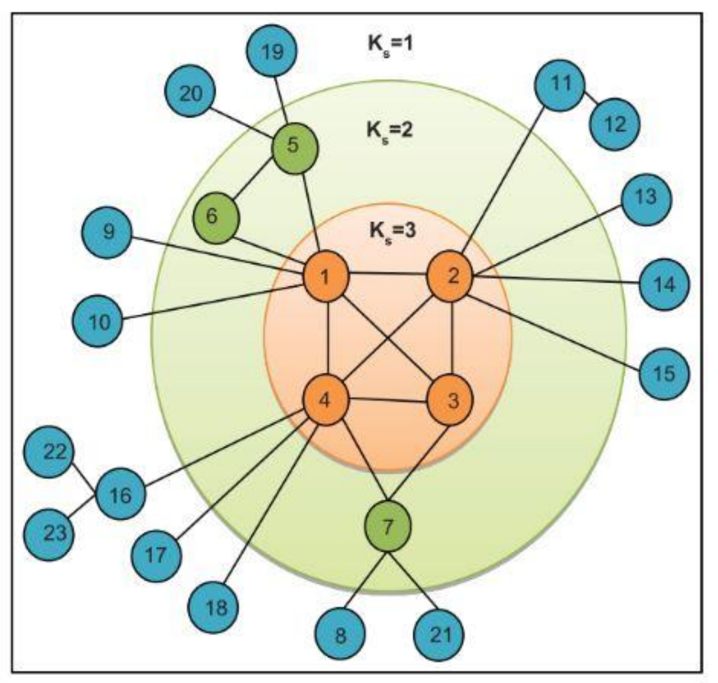
- A hybrid method which takes 3 factors into consideration:
  - Core Number of a node (i.e. diversity in different shells)
  - Degree of a node
  - Entropy (to calculate the dispersion of node v's friends in different cores.)

$$MCDWE(v) = \alpha Core(v) + \beta Degree(v) + \gamma Weighted\_Entropy(v)$$

$$\text{Entropy}(v) = -\sum_{i=0}^{Core_{max}} (p_i * \log_2 p_i)$$

$$p_i = \frac{\text{Count}(v\text{'s friends in core } i)}{\text{Degree}(v)}$$

$$Weighted\_Entropy(v) = -\sum_{i=0}^{Core_{max}} \frac{1}{(Core_{max} - Core_i + 1)} (p_i * \log_2 p_i)$$



Methods Implemented

Borda Count

- Single score by considering multiple ranking lists.

- Different ranked list considered :
  - Page-Rank
  - Degree Centrality
  - MCDWE rank

Position	RankingList1	RankingList2	RankingList3
1st Choice	A	C	D
2nd Choice	B	B	C
3rd Choice	C	D	B
4th Choice	D	A	A

Items	Borda Score
A	(1/1)+(1/4)+(1/4) = 1.5
B	(1/2)+(1/2)+(1/3) = 1.33
C	(1/3)+(1/1)+(1/2) = 1.83
D	(1/4)+(1/3)+(1/1) = 1.58

Methods Implemented

K-truss Decomposition

- Triangle based extension of a k-core decomposition.
- K-truss subgraph is the maximal subgraph where all edges belong to at least k-2 triangles

support(e) in H: #triangles e is in.
k-truss of G: largest subgraph H, each edge in H has support > k-2 in H.

Algo: **k-truss decompositon**
while(edges != 0)
  Loop: Number of edges reduced
    Loop: Edge
      If support(edge) < k-2
        Remove edge
        Edge score ← k - 1
  k ← k + 1

Node(v) = max(edge_score(e))

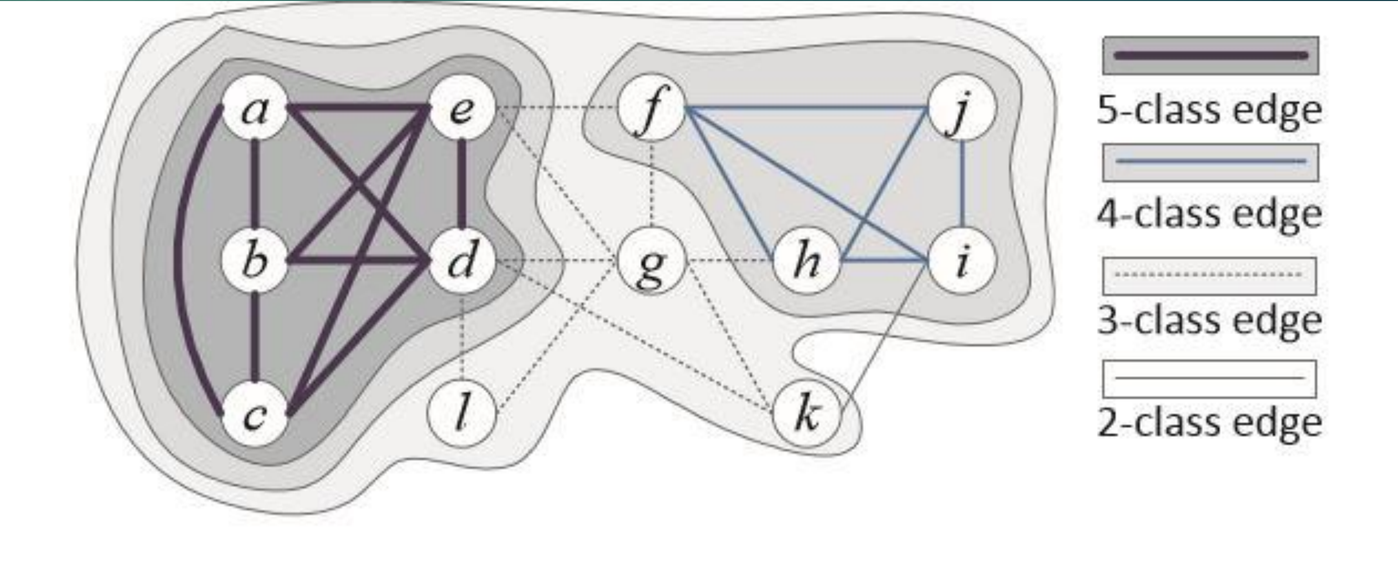


Figure 2: A graph G and the k-classes of G ($2 \leq k \leq 5$)

Evaluation Metrics

Methods used to evaluate ranking methods.

- **Overlap between ranked lists**
- **Average relative gain**
- **Average gain based on exposure**
- **Single/Multi Seed Simulation**

Evaluation Metrics

Overlap between ranked lists

- Given the set of top-k static influencers in two lists T and S.

$$O = \frac{|T \cap S|}{k}$$

k=100	Degree Centrality	Page Rank	K-truss	Broda Count	MCDWE Score
Degree Centrality	1	0.9	0.34	0.76	0.74
Page Rank	0.9	1	0.29	0.73	0.54
K-truss	0.34	0.29	1	0.4	0.27
Broda Count	0.76	0.73	0.4	1	0.58
MCDWE Score	0.74	0.54	0.27	0.58	1

k=200	Degree Centrality	Page Rank	K-truss	Broda Count	MCDWE Score
Degree Centrality	1	0.895	0.505	0.72	0.455
Page Rank	0.895	1	0.435	0.68	0.43
K-truss	0.505	0.435	1	0.525	0.445
Broda Count	0.72	0.706667	0.525	1	0.555
MCDWE Score	0.455	0.43	0.445	0.555	1

k=300	Degree Centrality	Page Rank	K-truss	Broda Count	MCDWE Score
Degree Centrality	1	0.893333333	0.52	0.73	0.436666667
Page Rank	0.89333	1	0.46	0.68	0.41
K-truss	0.52	0.46	1	0.543333333	0.5
Broda Count	0.73	0.68	0.5433	1	0.593333333
MCDWE Score	0.43667	0.41	0.5	0.593333	1

Evaluation Metrics

Average Relative Gain

- Impact of the retweet of a specific user on the final size of a cascade.

$$\mathcal{R}_u = \frac{1}{N_u} \sum_{i=1}^{N_u} \frac{n_i - k_i}{k_i}$$

Toy Network

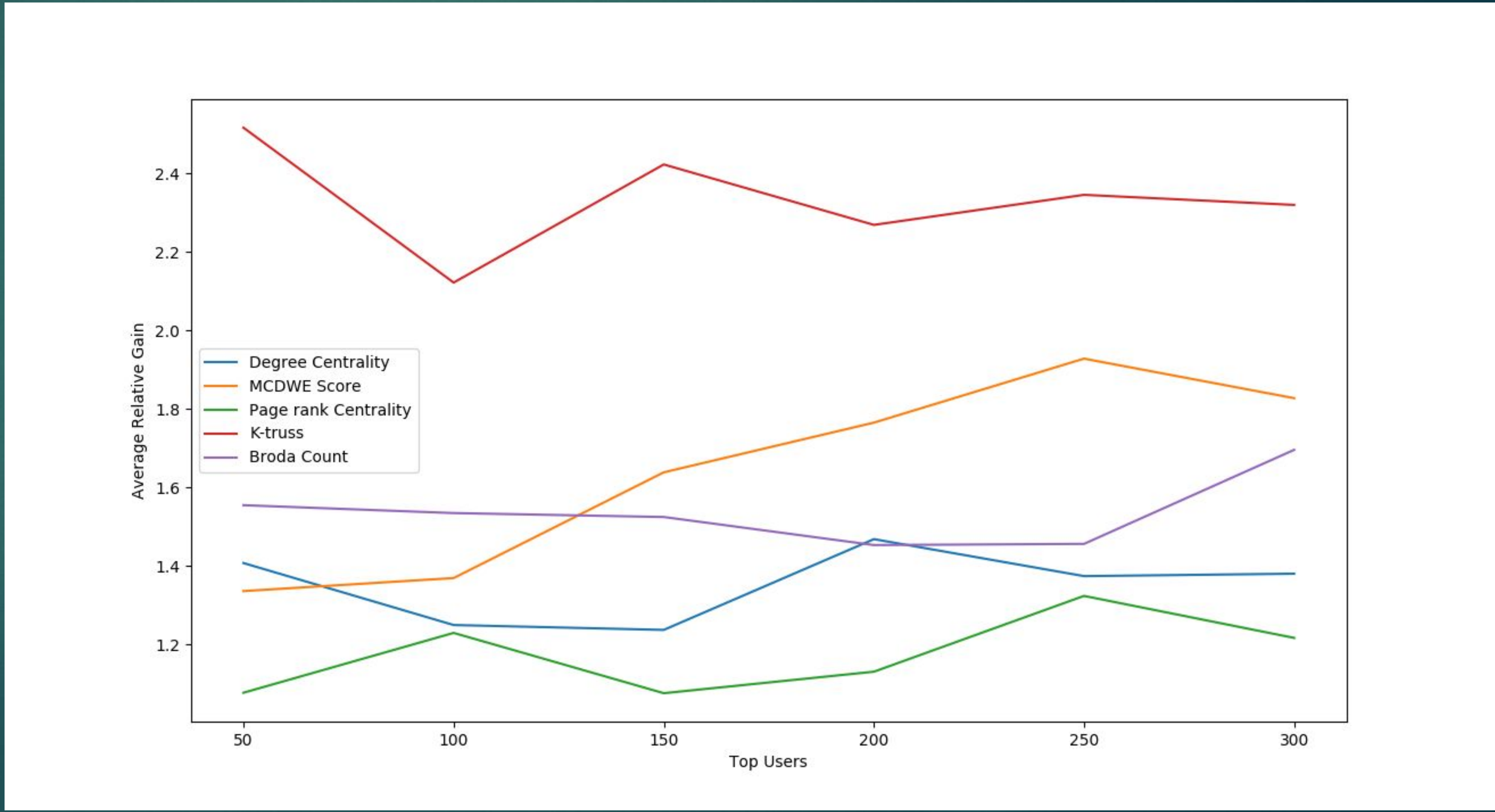
```
twt1 : B A D E G C F
twt2 : A B E
txt3 : C E A B

for node A

I(A)     = ((7-2)/2 + (3-1)/1 + (4-3)/3) / 3
         = 1.61

I(B)     = ((7-1)/1 + (4-4)/4) / 2
         = 3

I(E)     = ((7-4)/4 + (3-3)/3 + (4-2)/2) / 3
         = 0.583
```

Evaluation Metrics

Average Relative Gain

- Impact of the retweet of a specific user on the final size of a cascade.

$$\mathcal{R}_u = \frac{1}{N_u} \sum_{i=1}^{N_u} \frac{n_i - k_i}{k_i}$$

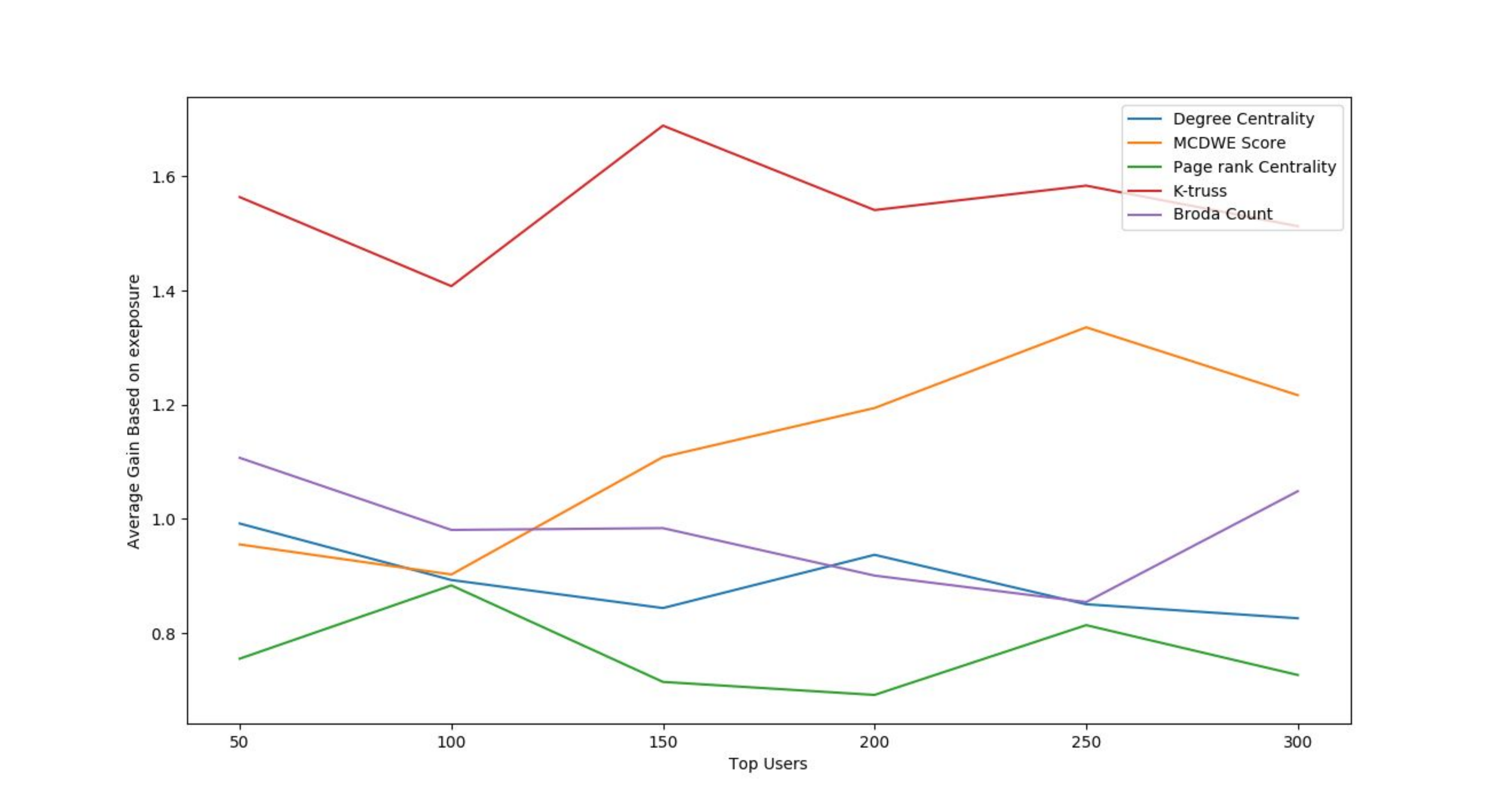


Evaluation Metrics

Average Gain Based on Exposure.

- For a user u in a cascade C of size n, this metric measures the number of re-tweeters after u that were newly exposed to C due to retweet by u if C is the i^{th} cascade in which u retweeted.

$$\mathcal{E}_u = \frac{1}{N_u} \sum_{i=1}^{N_u} \frac{a_i}{n_i}$$



Evaluation Metrics

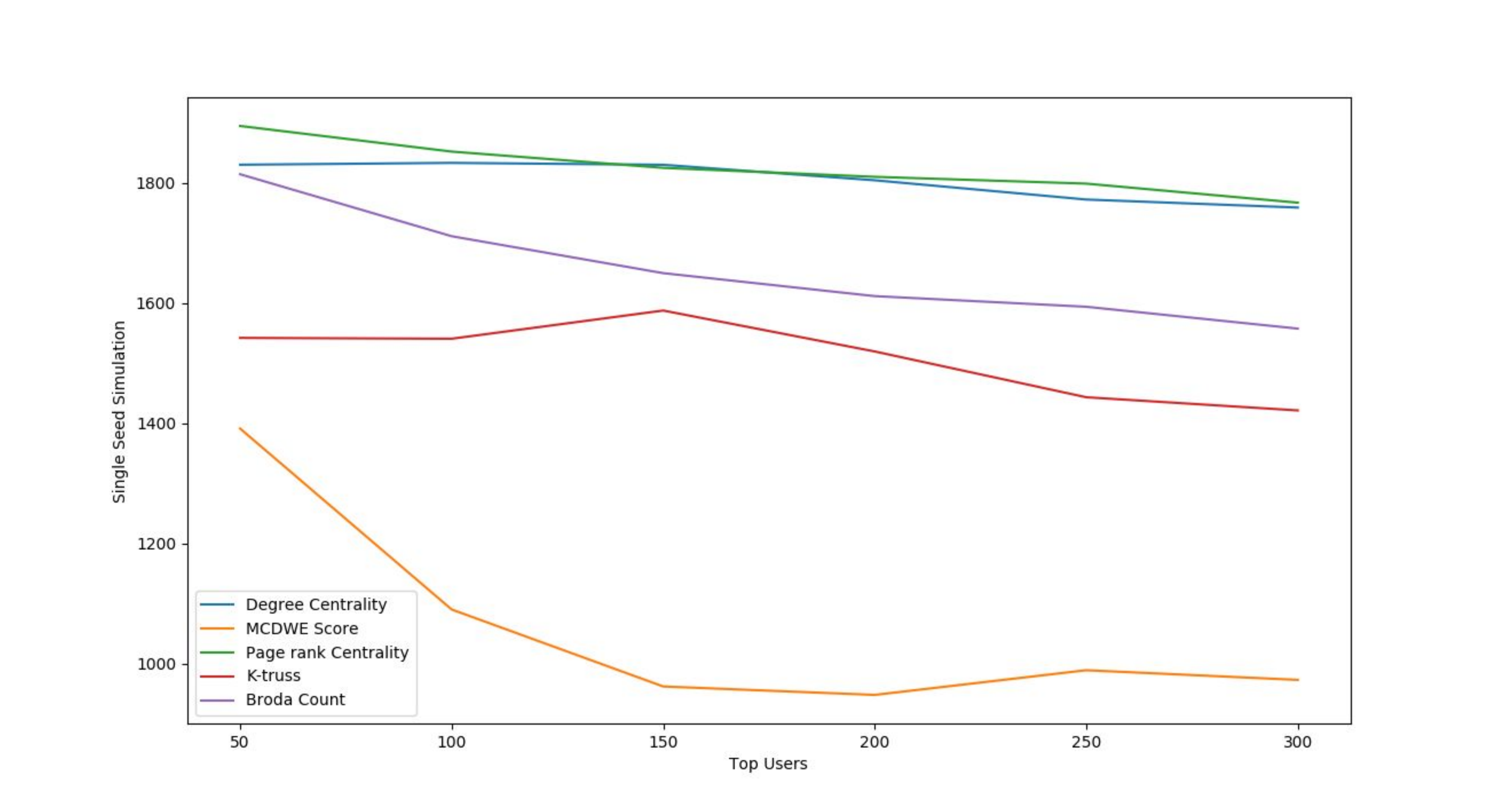
Single-seed/ Multi-seed Simulation

Single Seed Simulation:
- Computes theoretical number of infected nodes in network
- Seed set contains only one node

Multiple seed simulation:
- Seed set consists multiple nodes
- For each node simulate single cascade acting as a seed node

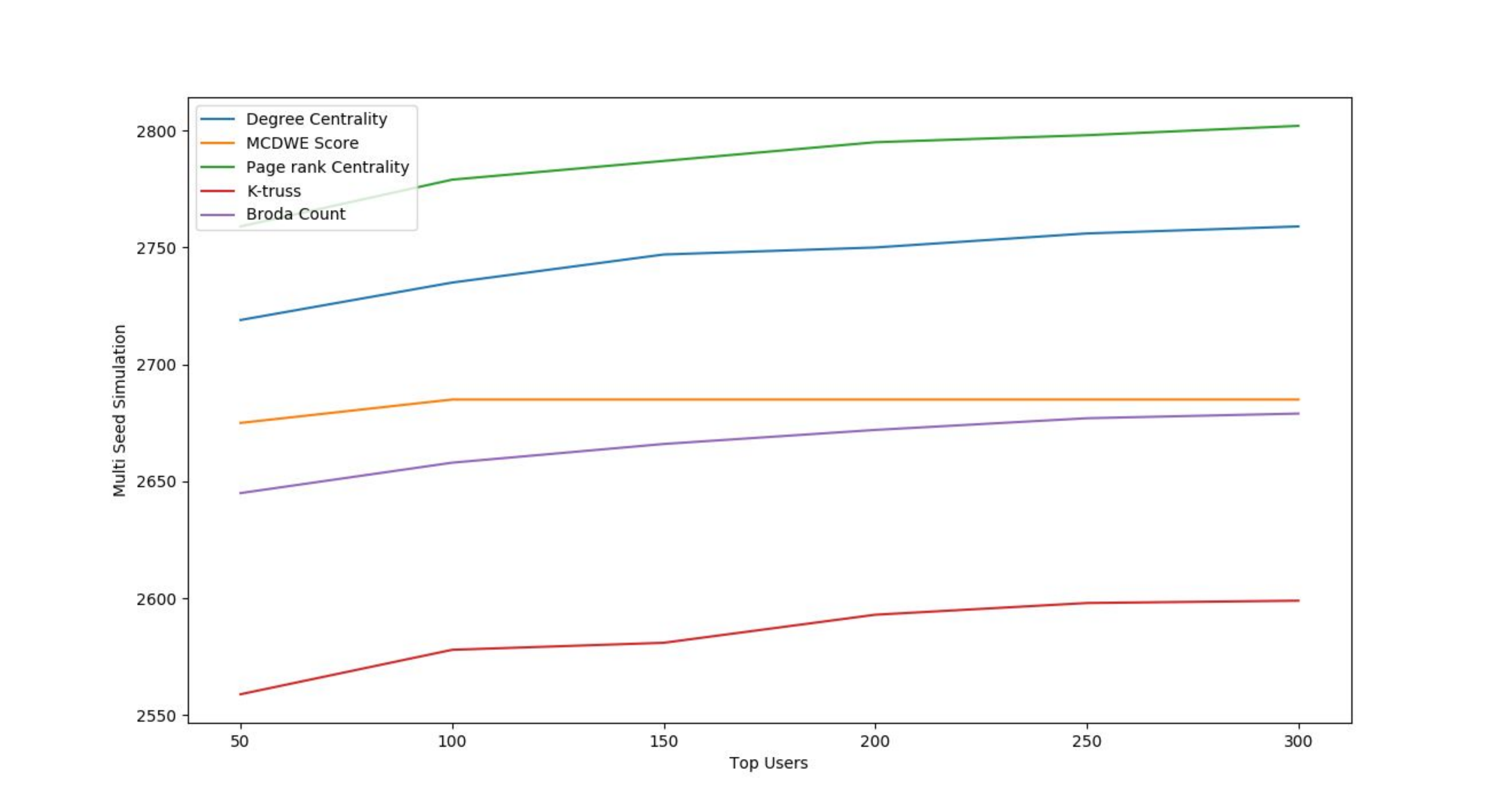
Evaluation Metrics

Single-seed Simulation



Evaluation Metrics

Multi-seed Simulation



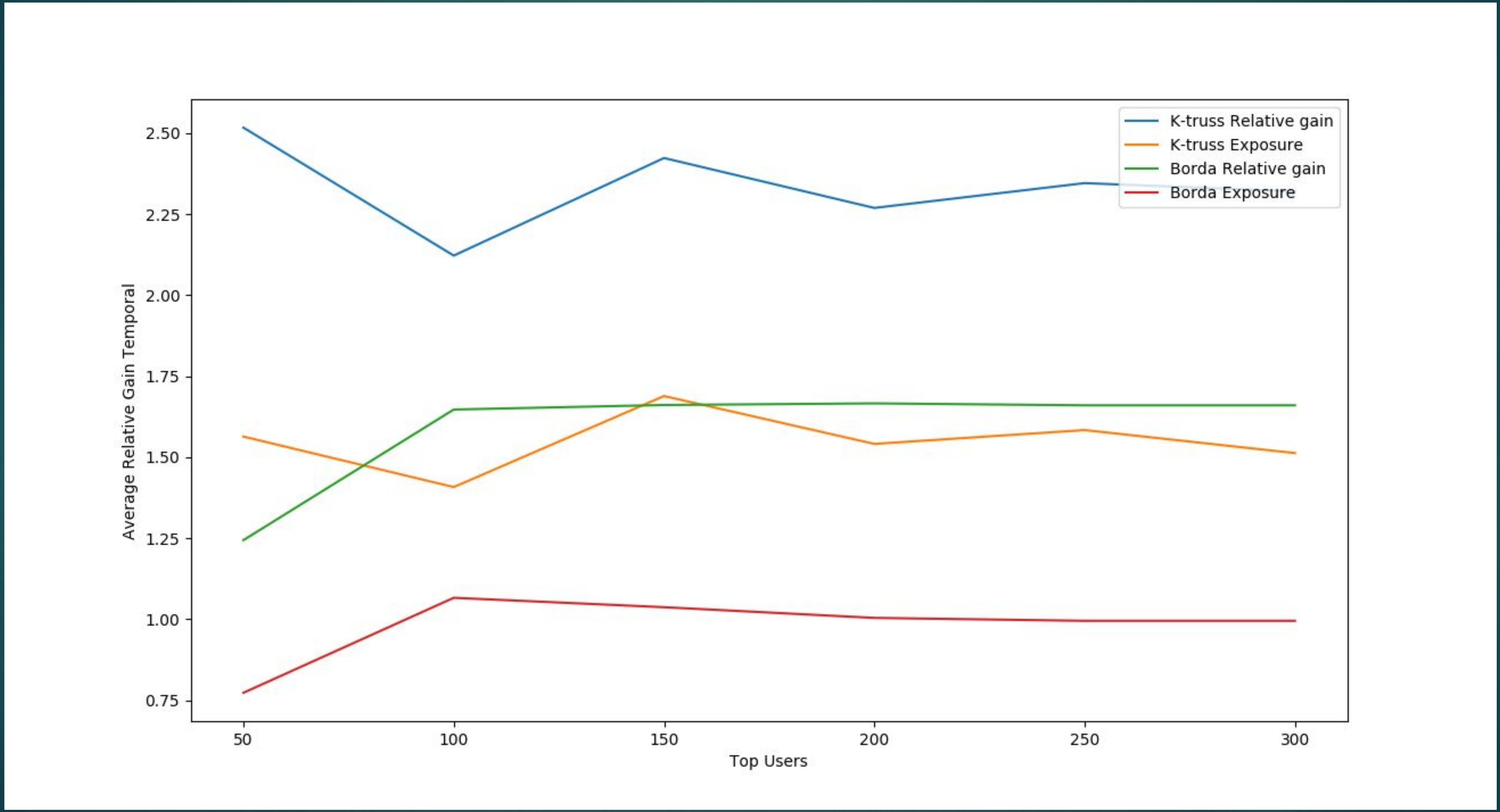
Temporal & Analysis(Part-2)

- Influential nodes discovery using temporal Data.
- Performance evaluation of temporal based method with structural based methods.

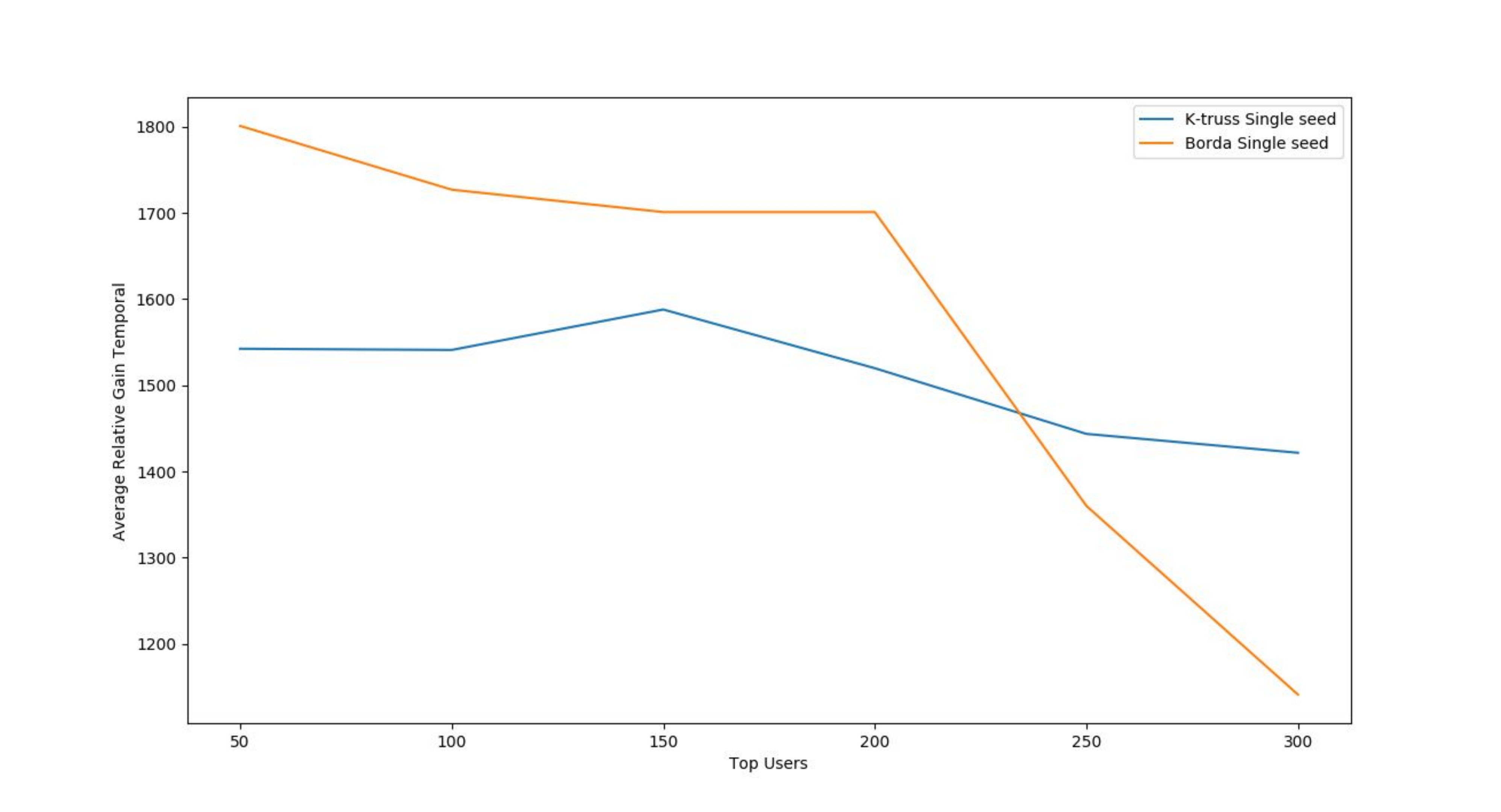
Why K-truss was not performing well for simulation

Degree Centrality	K-truss	Page-Rank	MCDWE
0.256	0.012	0.218	0.256
0.253	0.065	0.253	0.253
0.218	0.037	0.256	0.217
0.217	0.017	0.217	0.218
0.194	0.023	0.194	0.185
0.185	0.019	0.135	0.194
0.157	0.018	0.185	0.157
0.151	0.016	0.151	0.121
0.137	0.021	0.157	0.131
0.135	0.068	0.137	0.126

Why K-truss not in borda



Why K-truss not in borda



Temporal Influencer

Finding Influential nodes using the cascade information only.
Study evolution of inter retweet intervals of cascade.
Exploration pattern

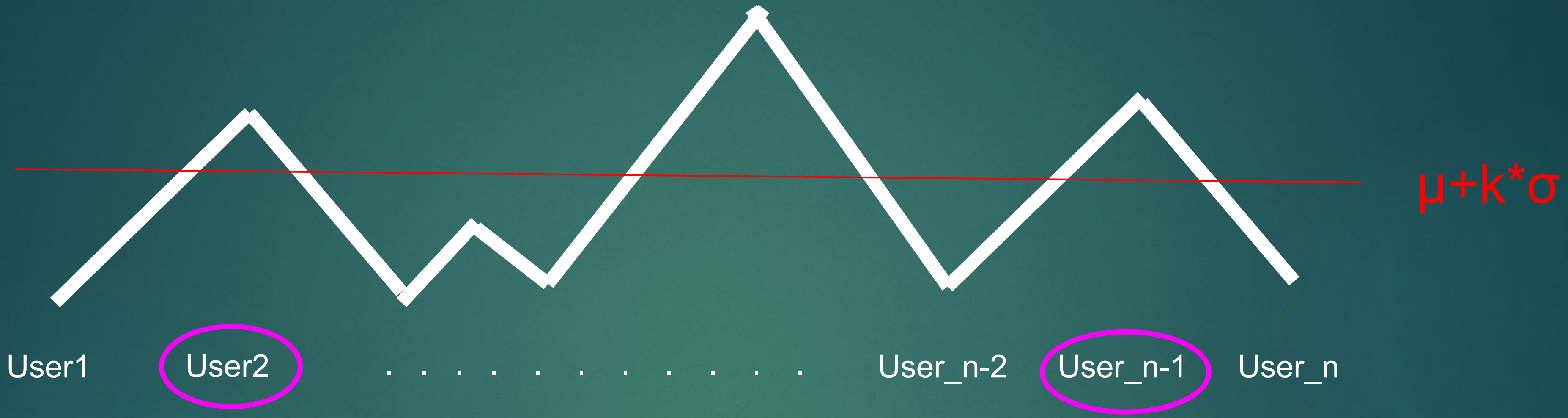
$$T^C = (T_0^C, T_1^C, T_2^C, \dots, T_{n-1}^C)$$

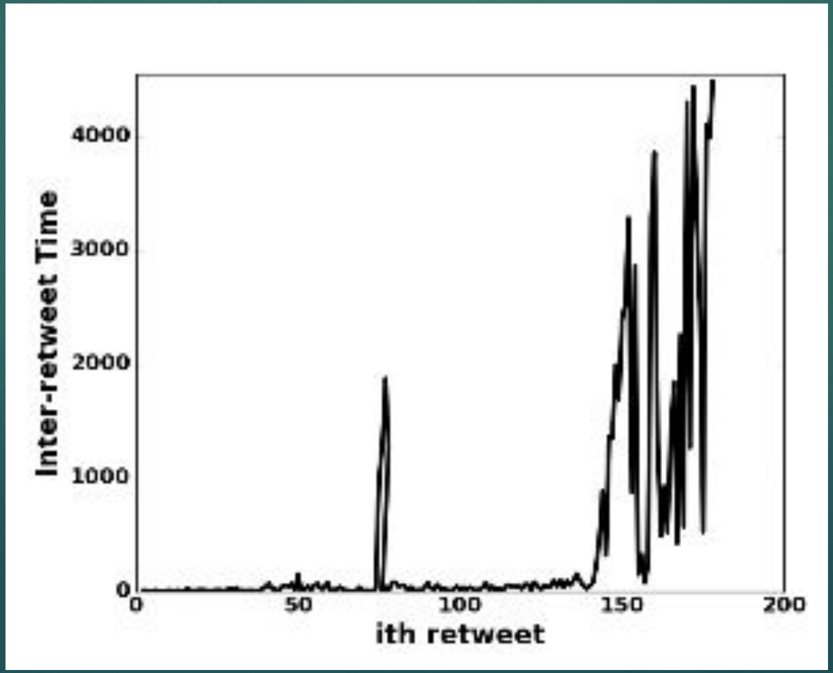
$$T_i^C = t_{i+1}^C - t_i^C$$

Peak Interval

```
Time interval cascade

XYZ : 10 11 20 15 19 5

mean     = 13.33
std      = 5.24

peak_interval    >= mean + (n * std)
                 >= 13.33 + 5.24
                 >= 18.58

peak_intervals = {20,19} -> {3,5}
```

Potential Temporal Influencers





How to Rank these influencers?

- Method A - Frequency of Retweets.
- Method B - Frequency of Retweets at peak time.
- Method C - Random selection of potential influencers.

Method A vs B

```
Retweets Cascades

A : u1 u2 u3 u4 u5
B : u3 u5 u2 u8 u10
C : u3 u6 u7 u1
D : u6 u4 u1 u2
E : u1 u2 u8

For user u2:
Frequency of Retweets                  - 4
Frequency of Retweets at peak time     - 2
```

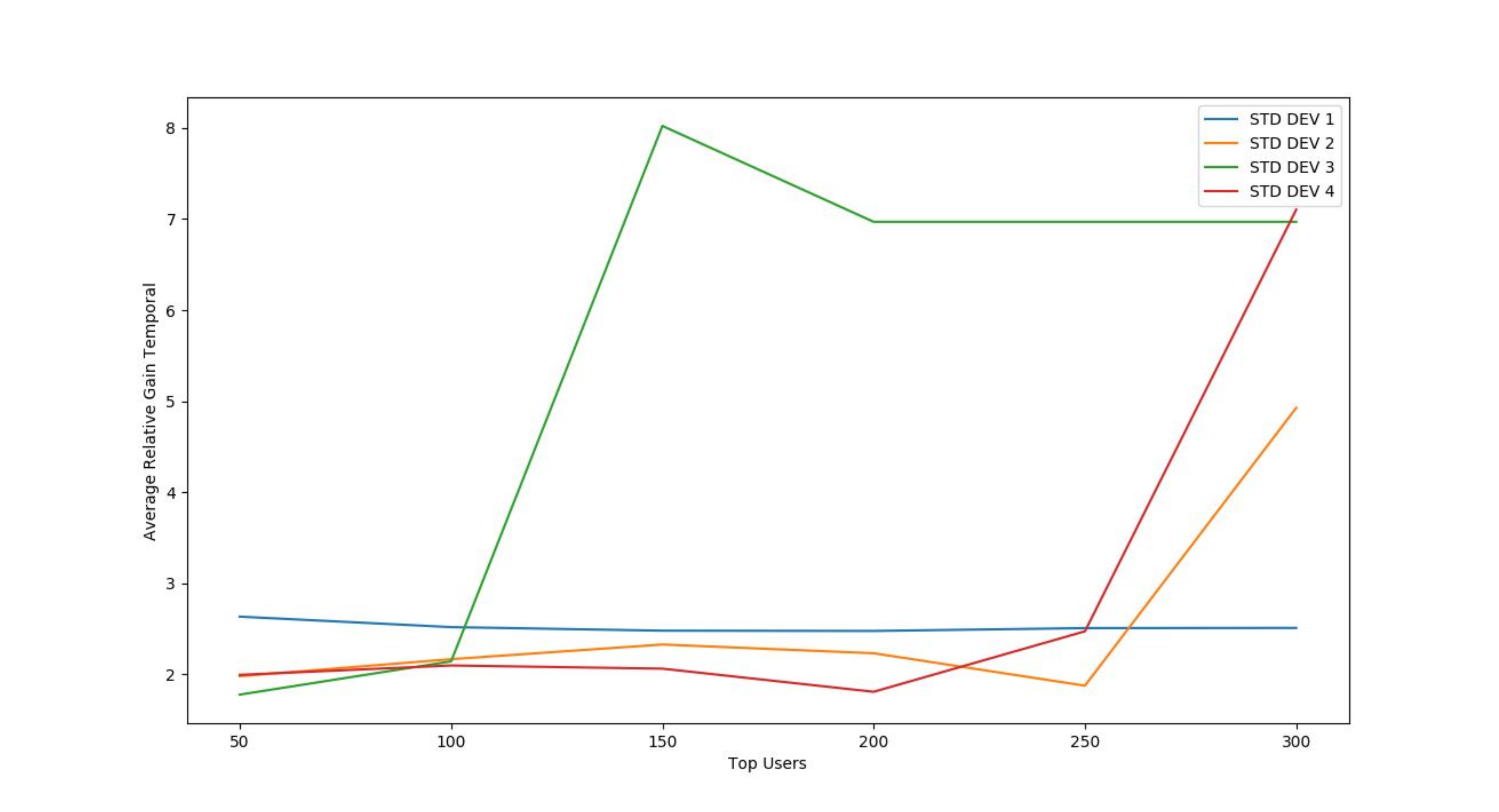
Results

- Comparison between different temporal methods.
- Comparison between different structural methods.
- Comparison between temporal and structural methods.
- Combining temporal and structural methods.

Best Structural Method : **K-truss**

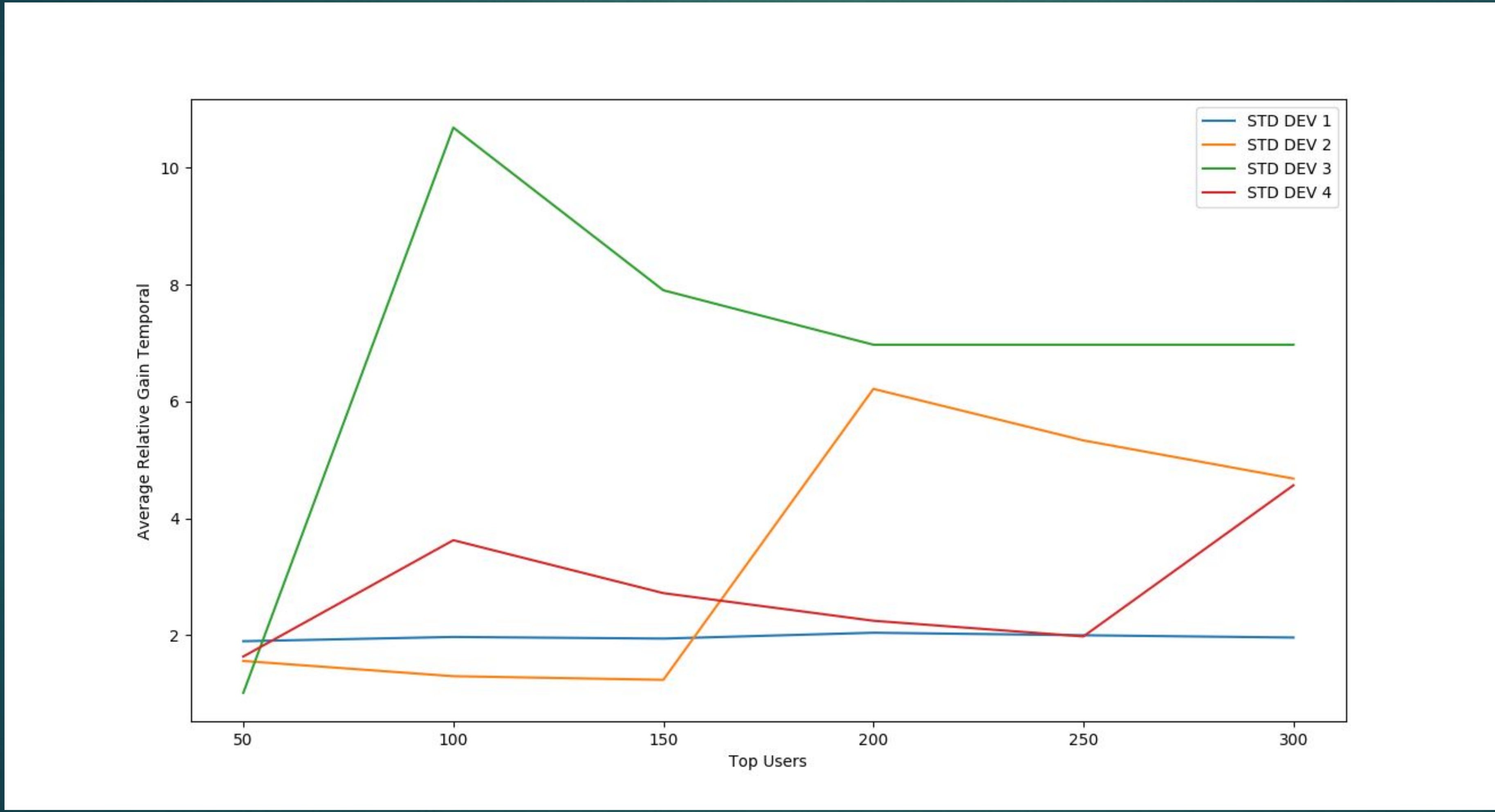
Best Temporal Method : **(mean + 3 * std)**

Average Relative gain Temporal



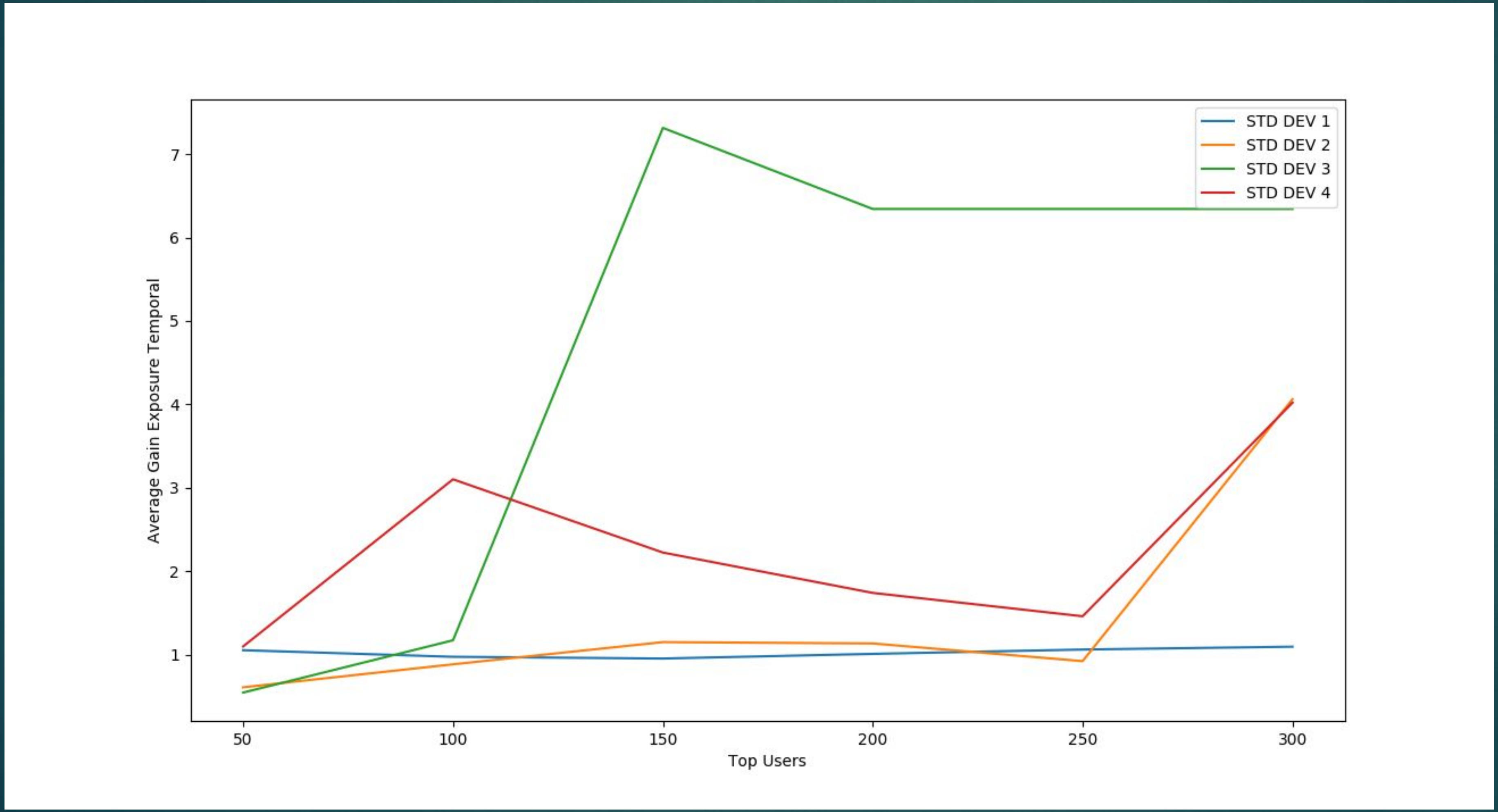
Method A

Average Relative gain Temporal



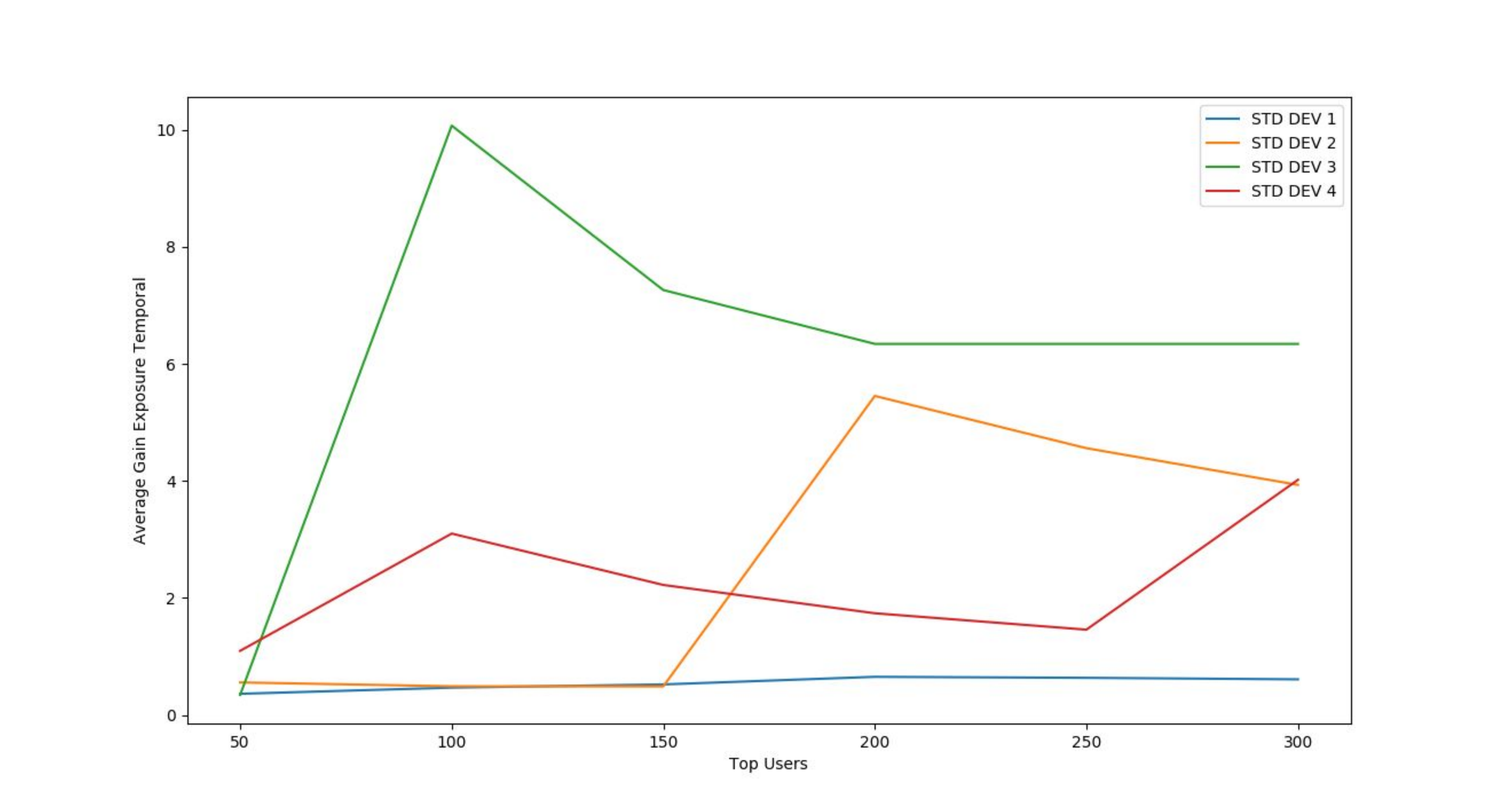
Method B

Average Gain based on exposure Temporal



Method A

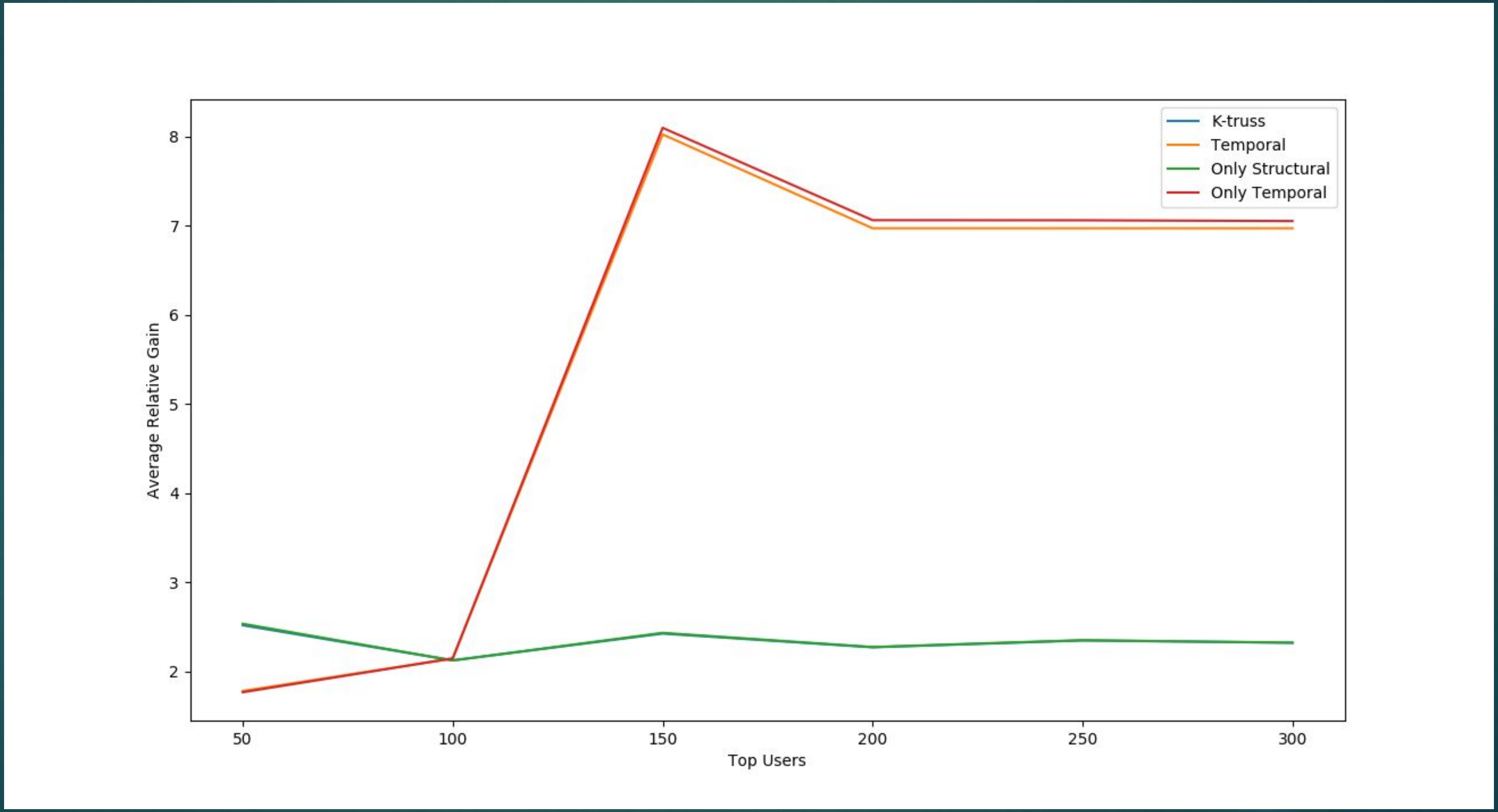
Average Gain based on exposure Temporal



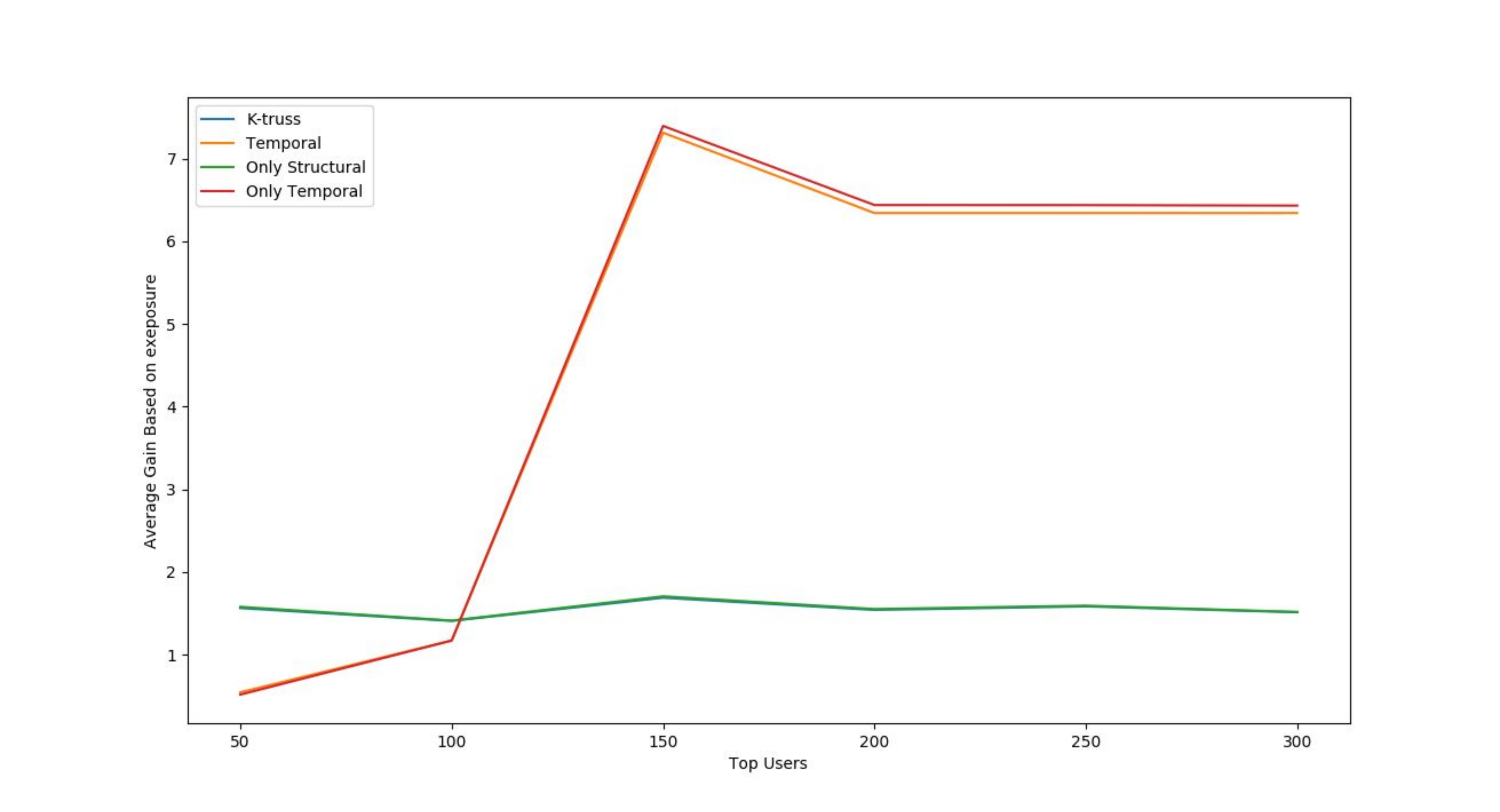
Method B

Average Relative Gain

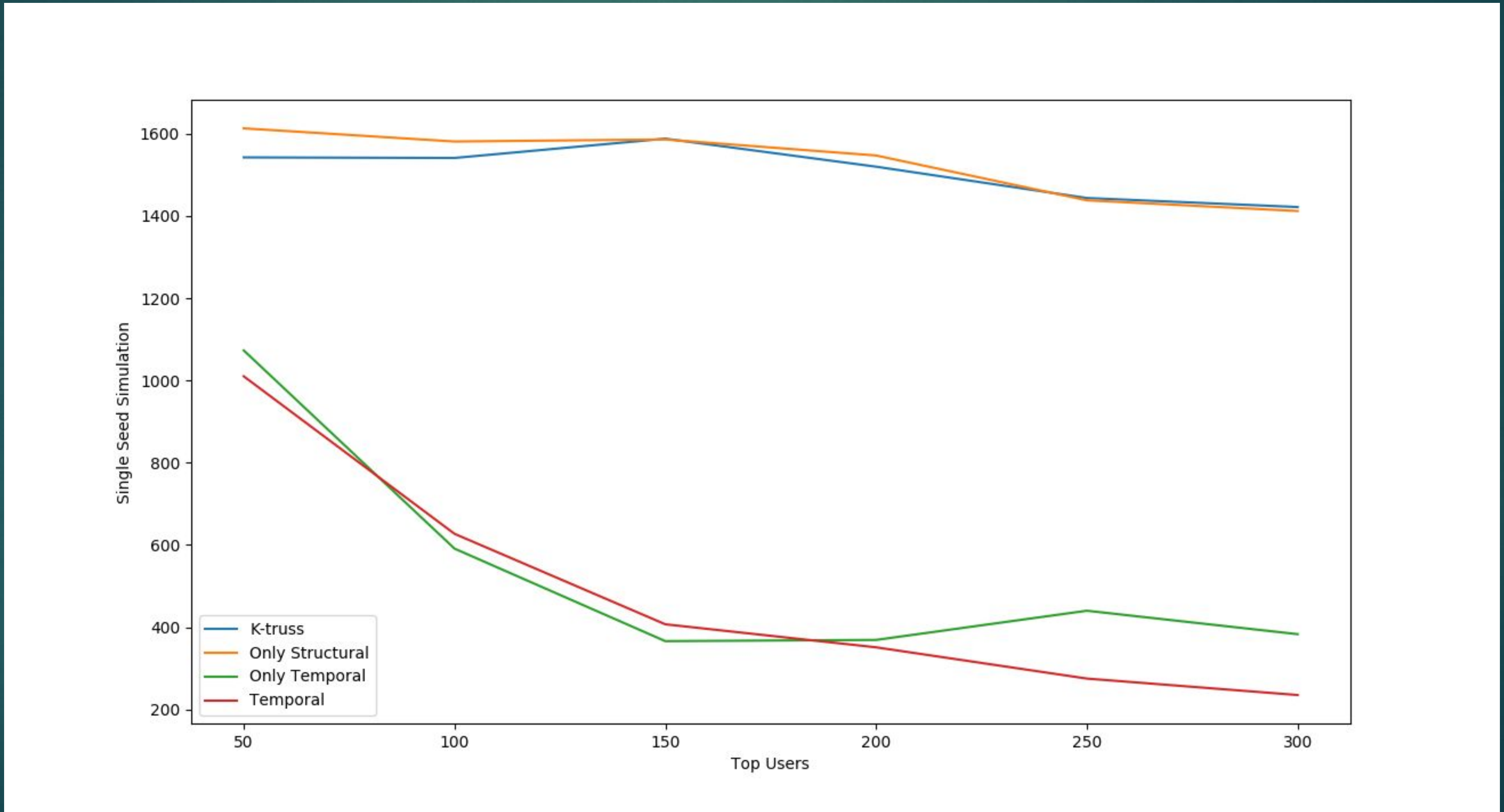
S - T => Only structural T - S => Only temporal



Average Gain based on exposure



Single Seed Simulation



Correlation between

Ground truth - Average relative gain

Y - value - Value of method A and B

Correlation 0.011

Conclusion

- Temporal retweet pattern of cascades are cheaply and readily available.
- Provides a very fast method to detect influencers.
- Find a better quality of influencers in terms of defined metrics, etc.

References

- Madotto, A and Liu, J. Super-Spreader Identification Using Meta-Centrality. Sci. Rep. 6,38994; doi: 10.1038/srep38994 (2016).
- Amir Sheikhahmadi et al. Identification of multi-spreader users in social networks for viral marketing.
- Malliaros, F. D. et al. Locating influential nodes in complex networks. Sci. Rep. 6, 19307; doi: 10.1038/srep19307 (2016).
- Bhowmick A. [Identification of influential users in the network using temporal patterns of(re)tweet cascades combined with network topology][Paper January 2018]
- Image references (Google Images)

Thank You!