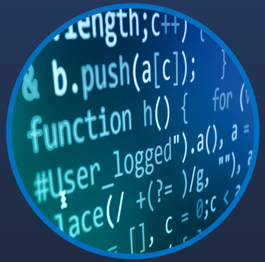
Big Data Era

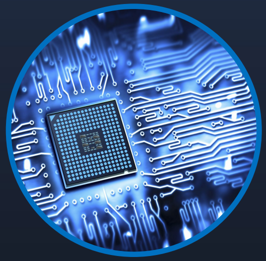
https://vimeo.com/102998774

The big problem: Scalability

Visualization



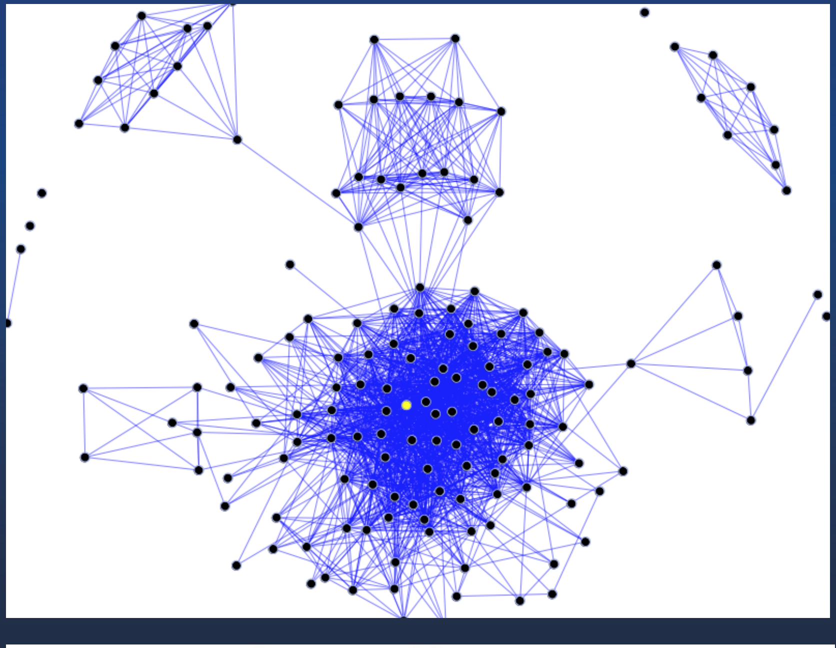
Algorithm

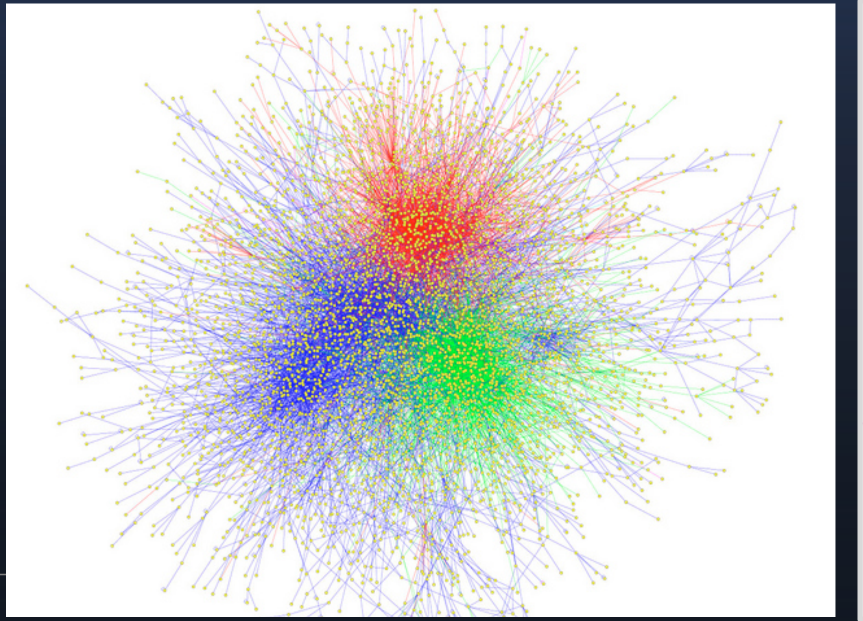


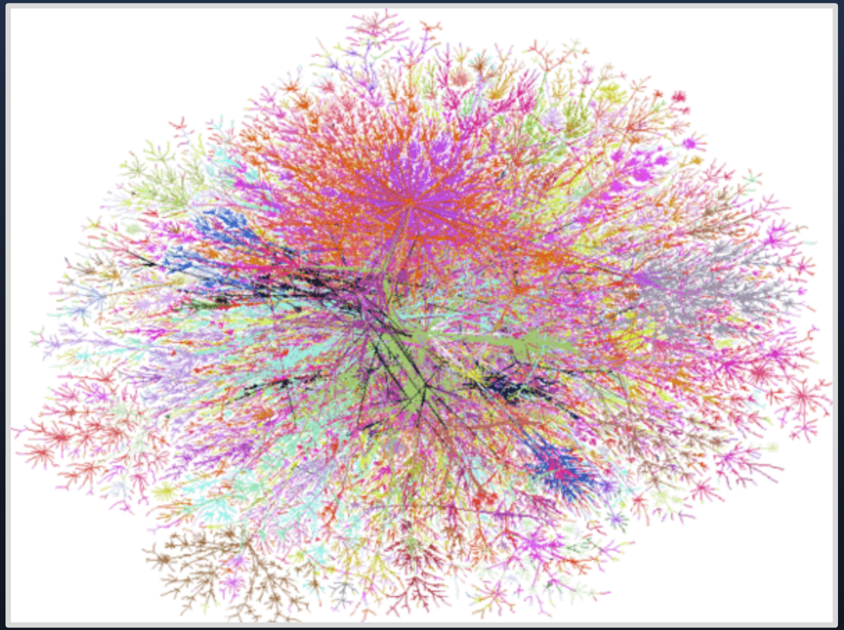
Hardware

The big problem: Scalability

- Visualization
- Algorithm
- Hardware



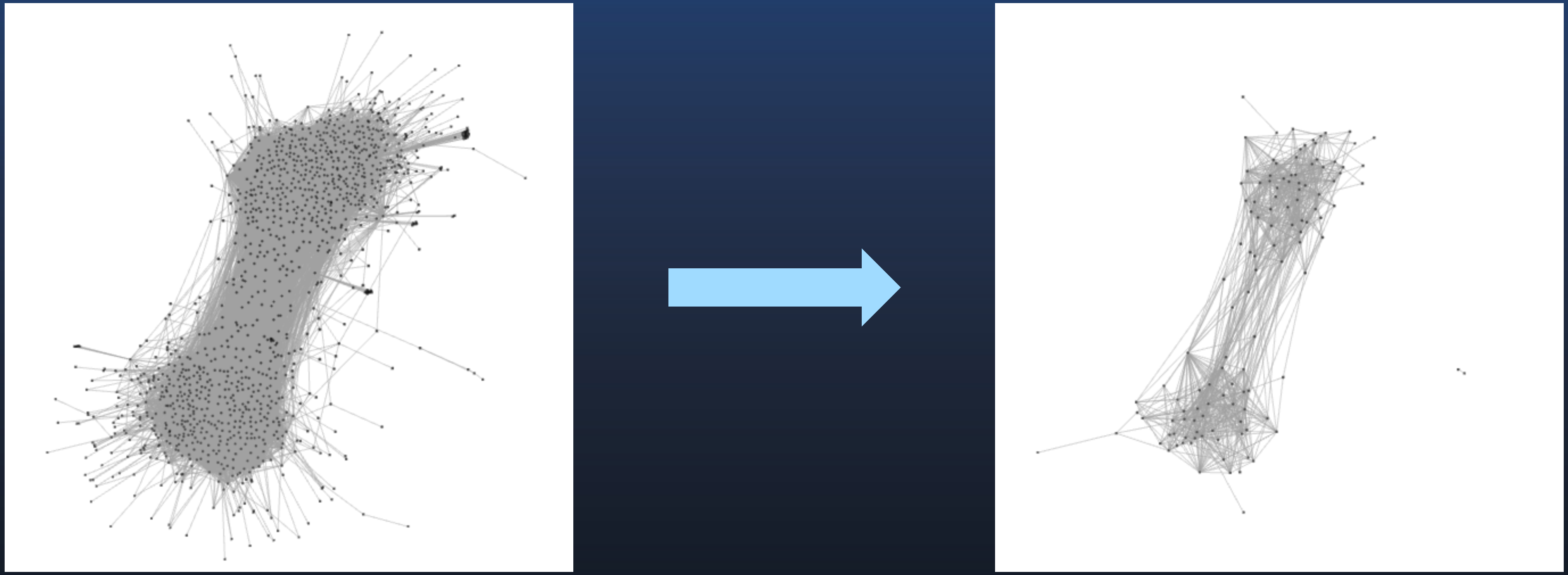




https://upload.wikimedia.org/wikipedia/commons/0/05/Sna_large.png
https://upload.wikimedia.org/wikipedia/commons/9/9b/Social_Network_Analysis_
https://c1.staticflickr.com/5/4033/4520018121_6dd39e8d7e_z.jpg
https://c1.staticflickr.com/1/1/916142_ddc2fd0140.jpg

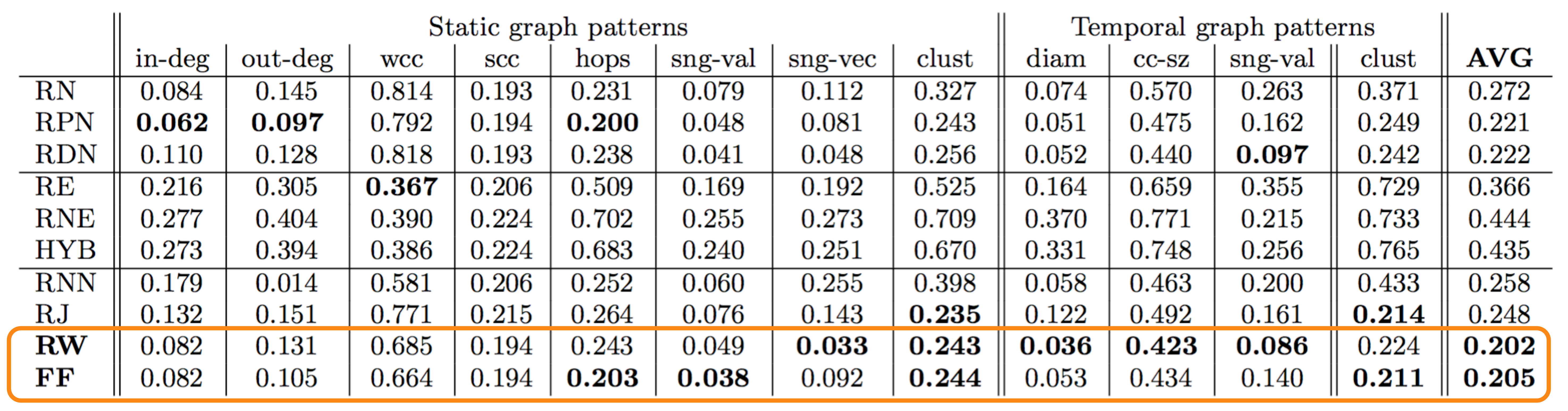
Graph Sampling

- Randomly pick nodes /edges to construct a subgraph that represents the original unfiltered graph:



Which sampling strategy to use?

Graph Sampling Evaluation

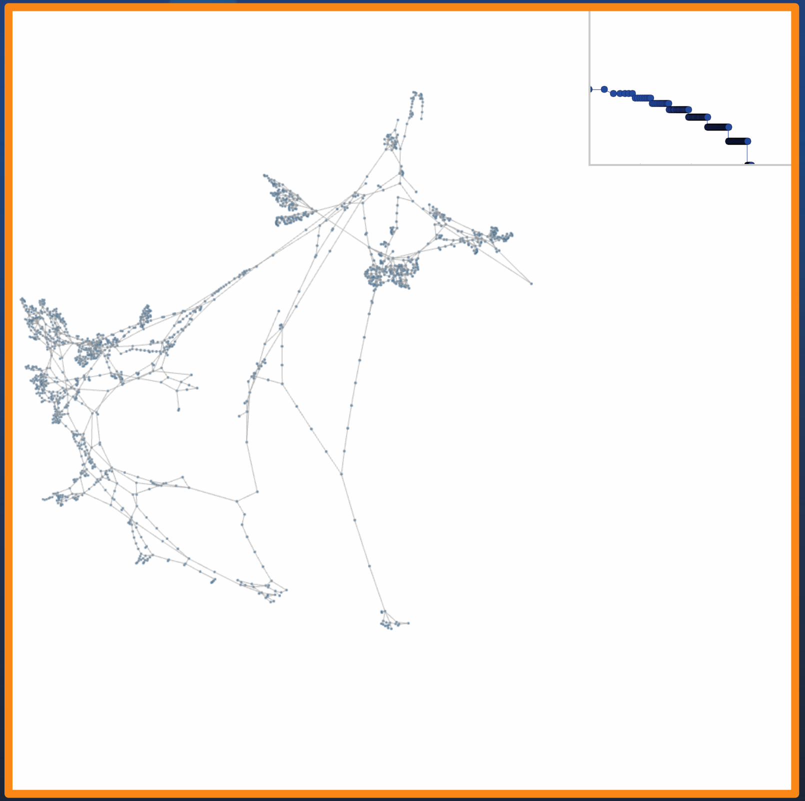


	Static graph patterns								Temporal graph patterns				
	in-deg	out-deg	wcc	scc	hops	sng-val	sng-vec	clust	diam	cc-sz	sng-val	clust	**AVG**
RN	0.084	0.145	0.814	0.193	0.231	0.079	0.112	0.327	0.074	0.570	0.263	0.371	0.272
RPN	**0.062**	**0.097**	0.792	0.194	**0.200**	0.048	0.081	0.243	0.051	0.475	0.162	0.249	0.221
RDN	0.110	0.128	0.818	0.193	0.238	0.041	0.048	0.256	0.052	0.440	**0.097**	0.242	0.222
RE	0.216	0.305	**0.367**	0.206	0.509	0.169	0.192	0.525	0.164	0.659	0.355	0.729	0.366
RNE	0.277	0.404	0.390	0.224	0.702	0.255	0.273	0.709	0.370	0.771	0.215	0.733	0.444
HYB	0.273	0.394	0.386	0.224	0.683	0.240	0.251	0.670	0.331	0.748	0.256	0.765	0.435
RNN	0.179	0.014	0.581	0.206	0.252	0.060	0.255	0.398	0.058	0.463	0.200	0.433	0.258
RJ	0.132	0.151	0.771	0.215	0.264	0.076	0.143	**0.235**	0.122	0.492	0.161	**0.214**	0.248
RW	0.082	0.131	0.685	0.194	0.243	0.049	**0.033**	**0.243**	**0.036**	**0.423**	**0.086**	0.224	**0.202**
FF	0.082	0.105	0.664	0.194	**0.203**	**0.038**	0.092	**0.244**	0.053	0.434	0.140	**0.211**	**0.205**

Random Walk (RW) v.s. Forest Fire (FF)

[Leskovec and Faloutsos, KDD 2006]

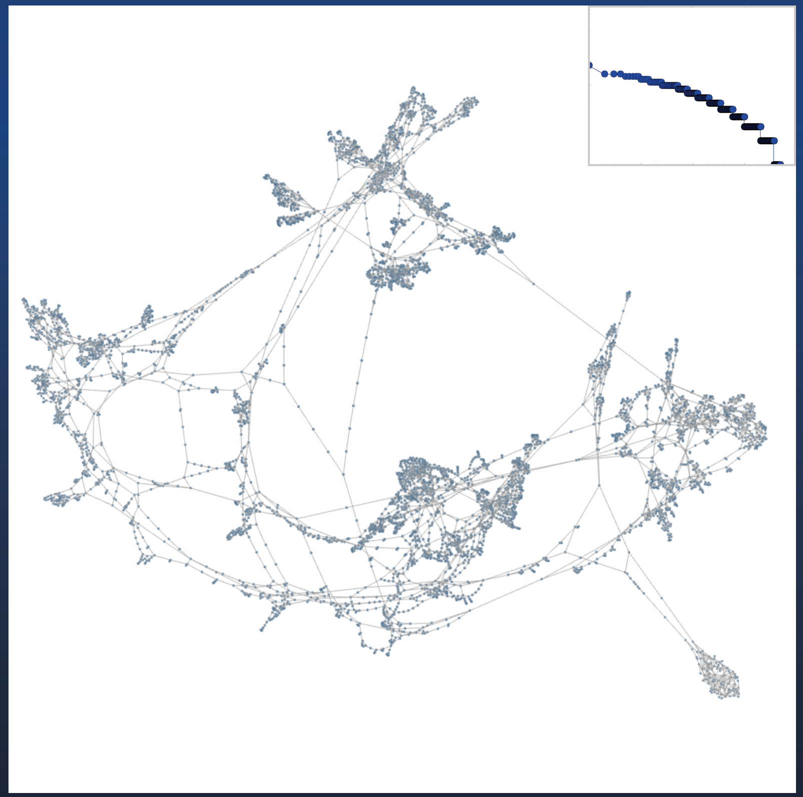
Graph Sampling Evaluation in Visualization



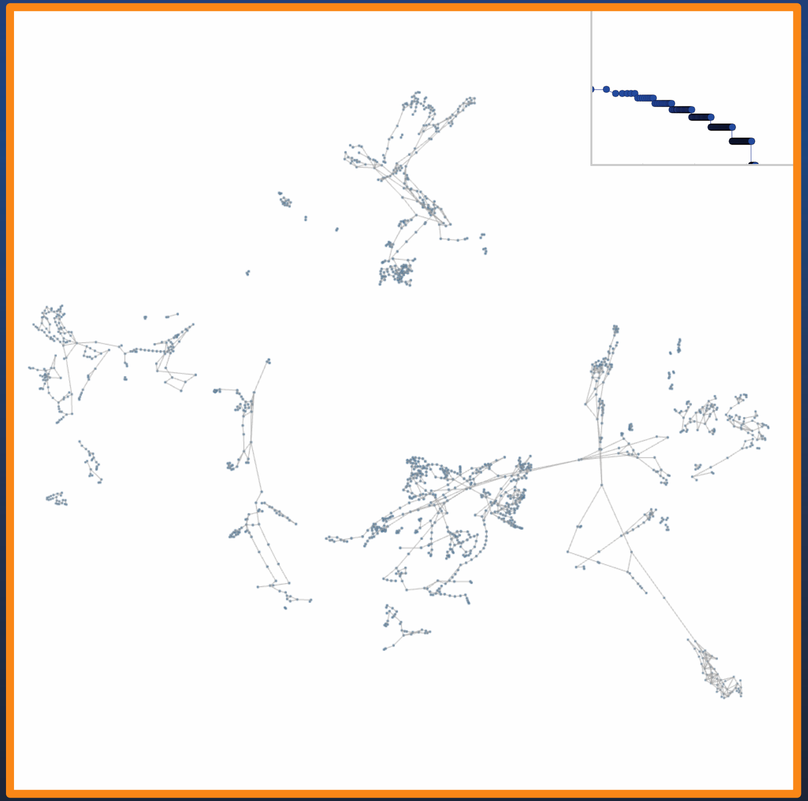
Random Walk (RW)

Avg. node degree: 2.4

Power-law degree distribution



Original Graph



Forest Fire (FF)

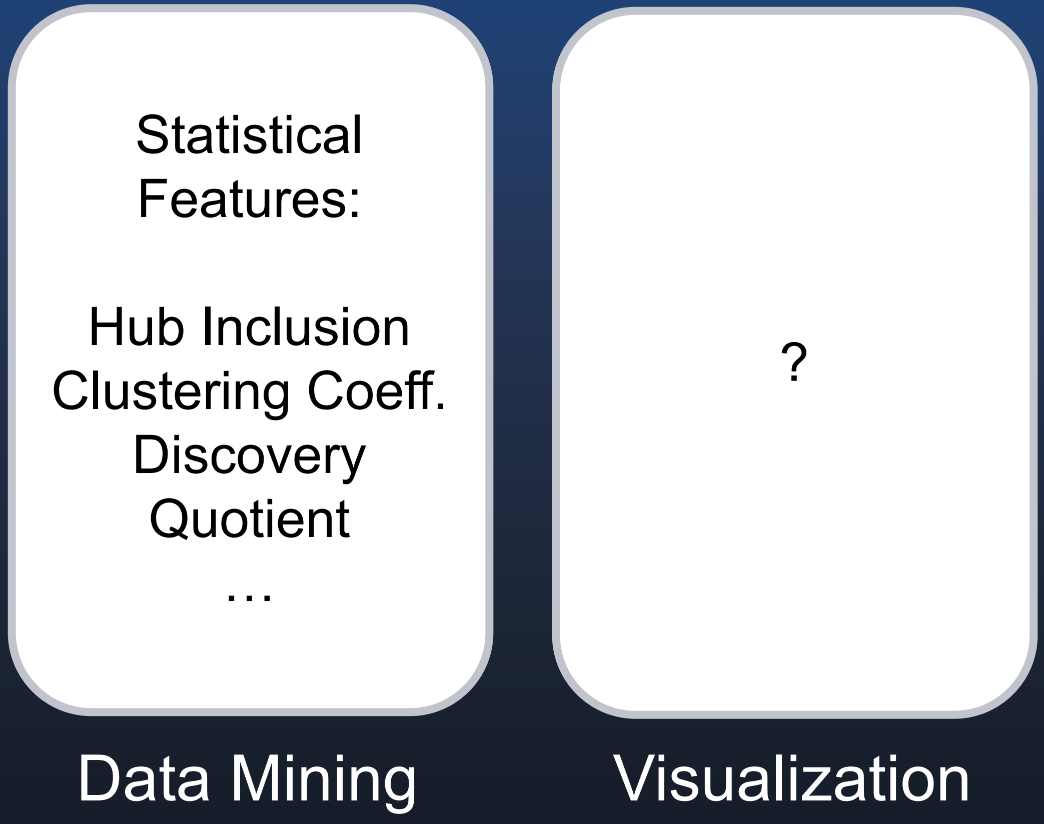
Avg. node degree: 2.4

Power-law degree distribution

Distinct Visual Result!

Graph Sampling Evaluation in Visualization

Similarity Measurements



Statistical Features:

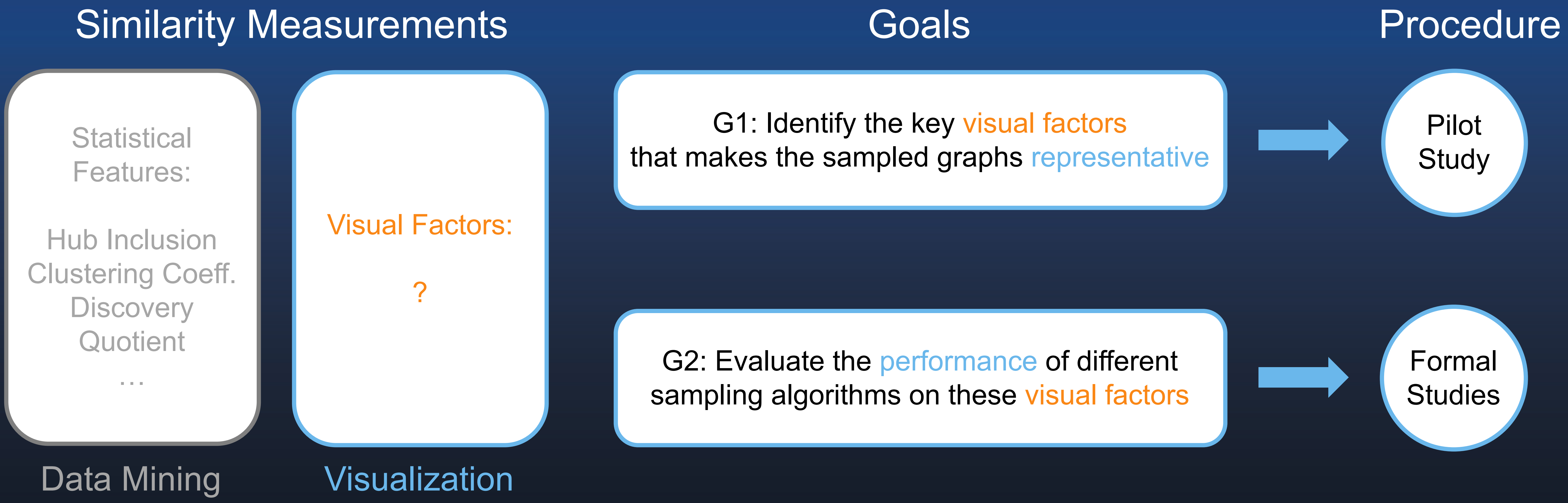
Hub Inclusion
Clustering Coeff.
Discovery Quotient
…

?

Data Mining

Visualization

Graph Sampling Evaluation in Visualization



Similarity Measurements

Statistical Features:

Hub Inclusion
Clustering Coeff.
Discovery Quotient
…

Data Mining

Visual Factors:

?

Visualization

Goals

G1: Identify the key visual factors that makes the sampled graphs representative

G2: Evaluate the performance of different sampling algorithms on these visual factors

Procedure

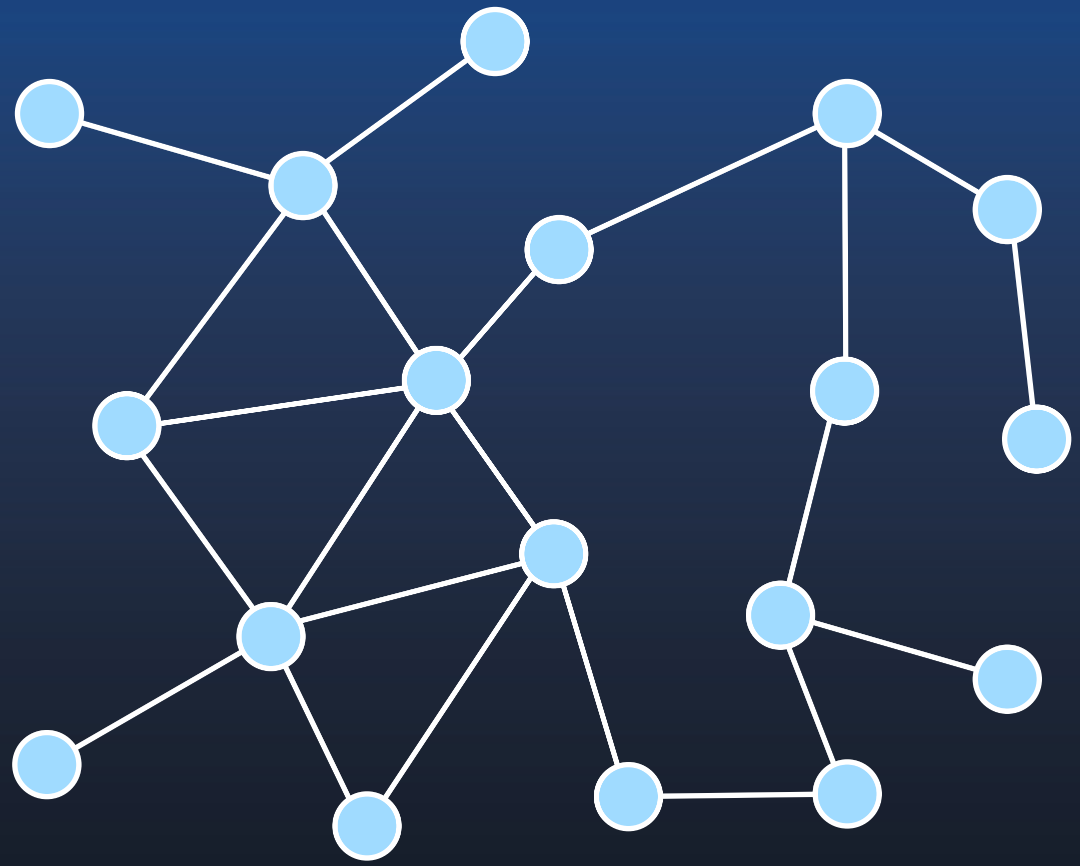
Pilot Study

Formal Studies

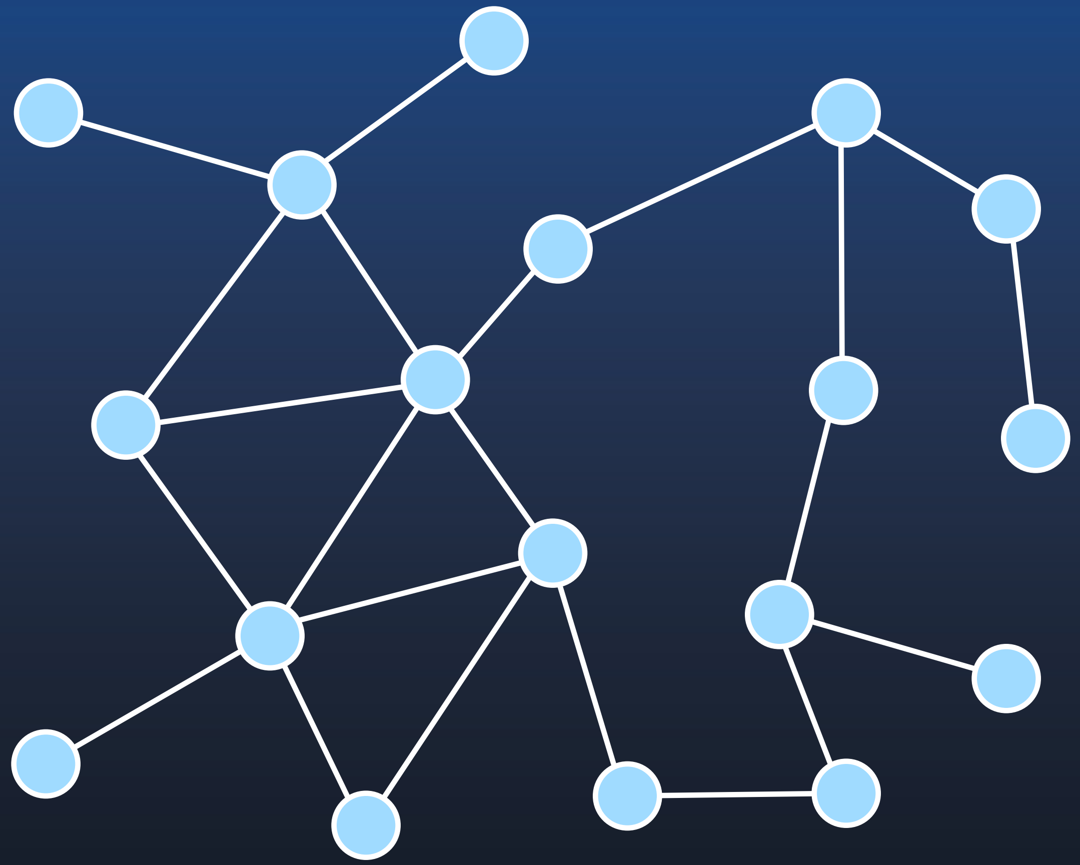
Outline

- Selected Sampling Methods
- Pilot Study
- Formal Studies
  - Perception of High Degree Nodes
  - Perception of Cluster Quality
  - Perception of Coverage Area

Node-Based Sampling

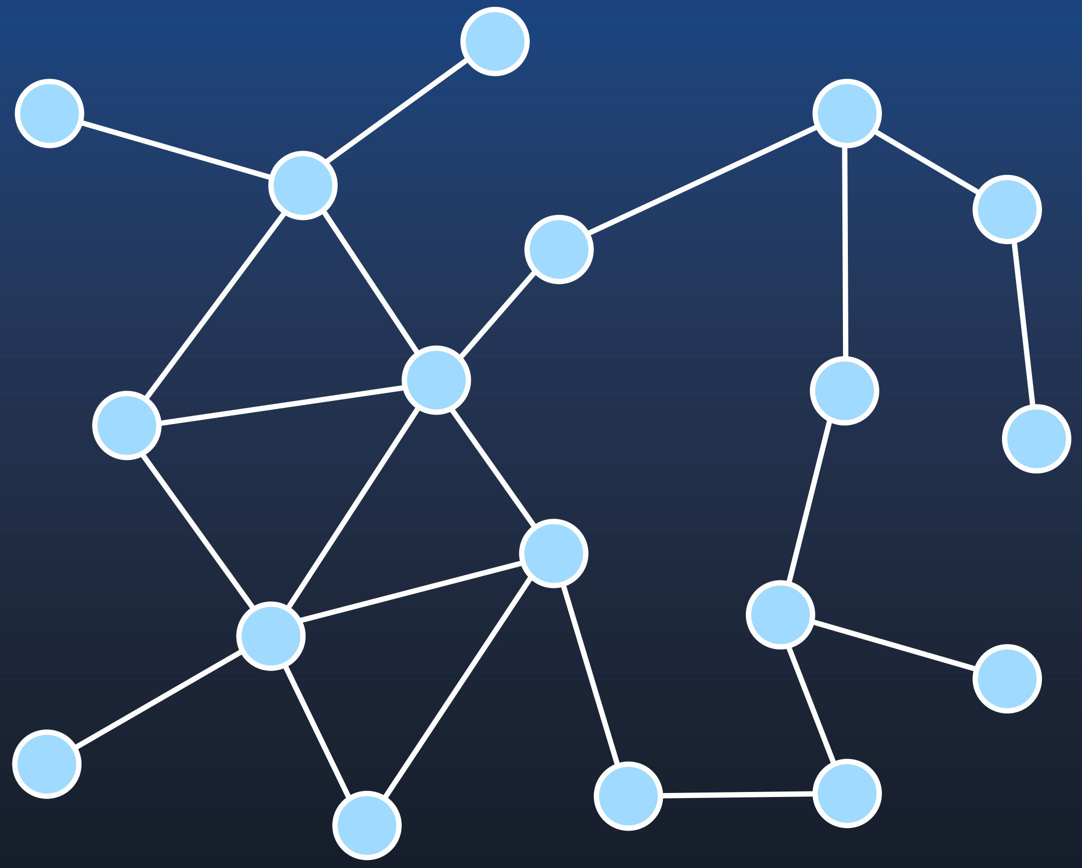


Original Graph

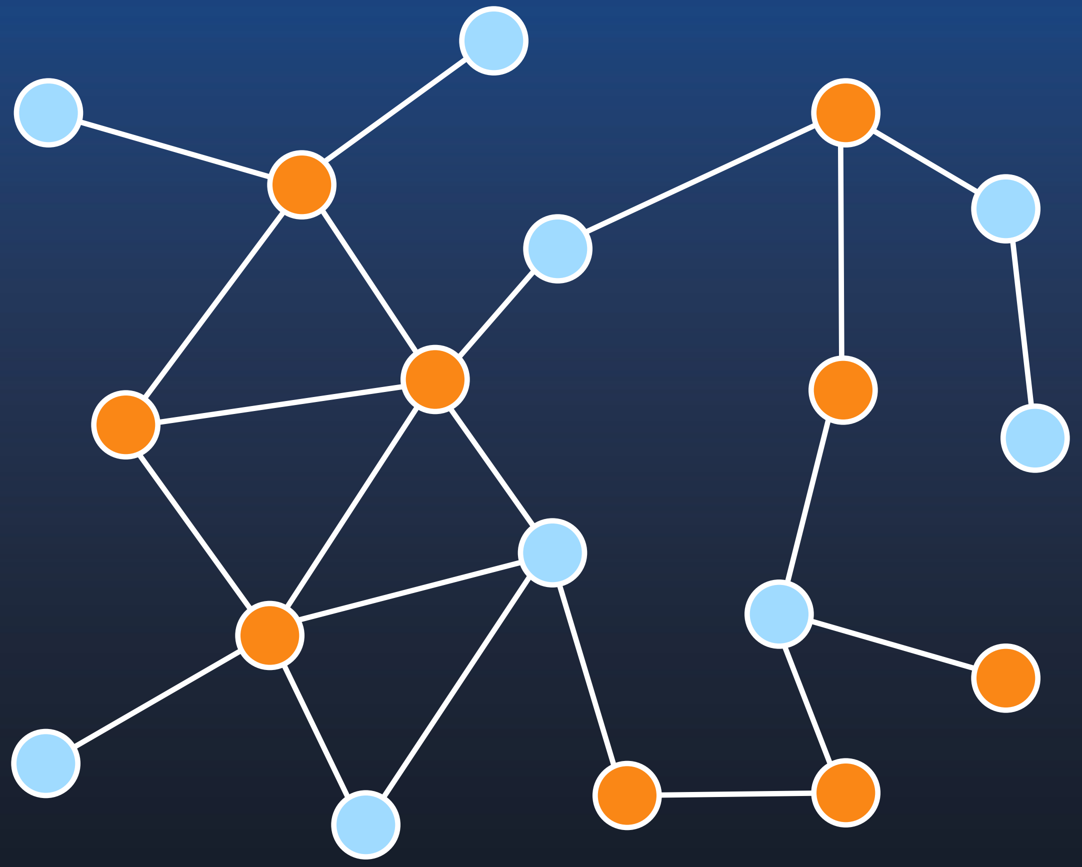


Random Node Sampling

Node-Based Sampling

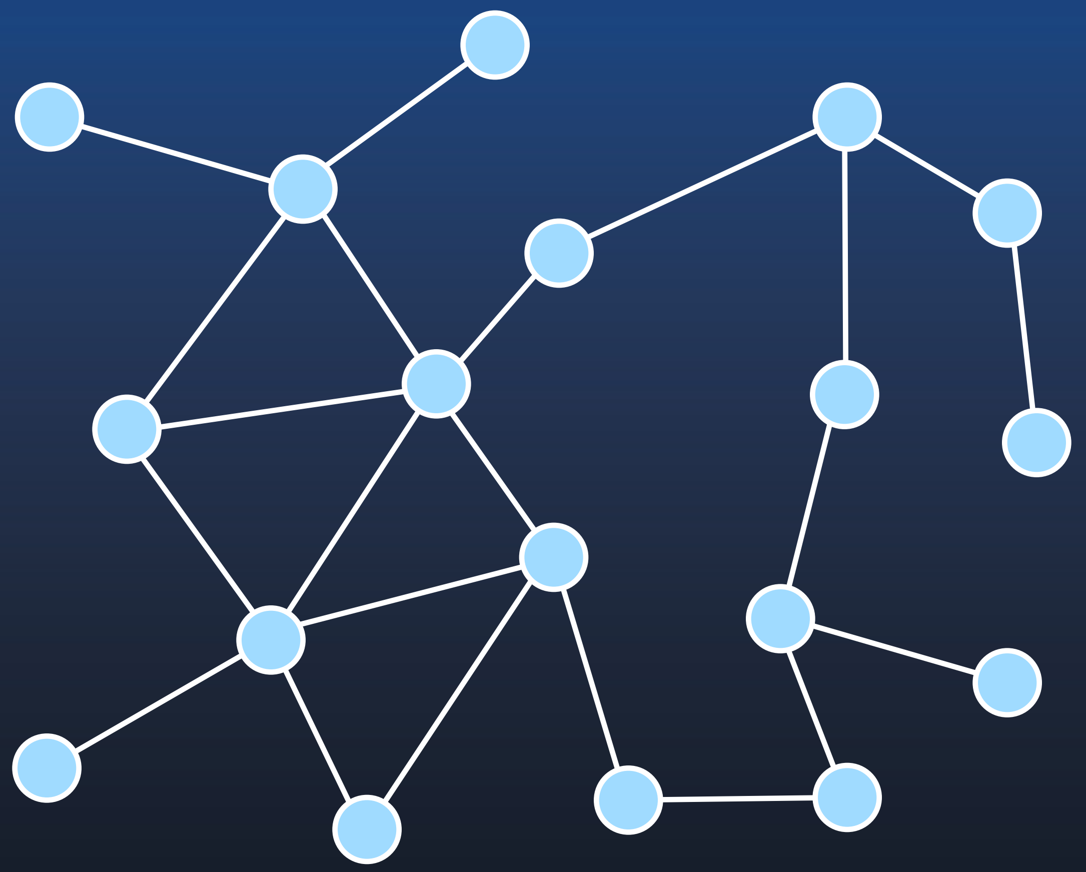


Original Graph

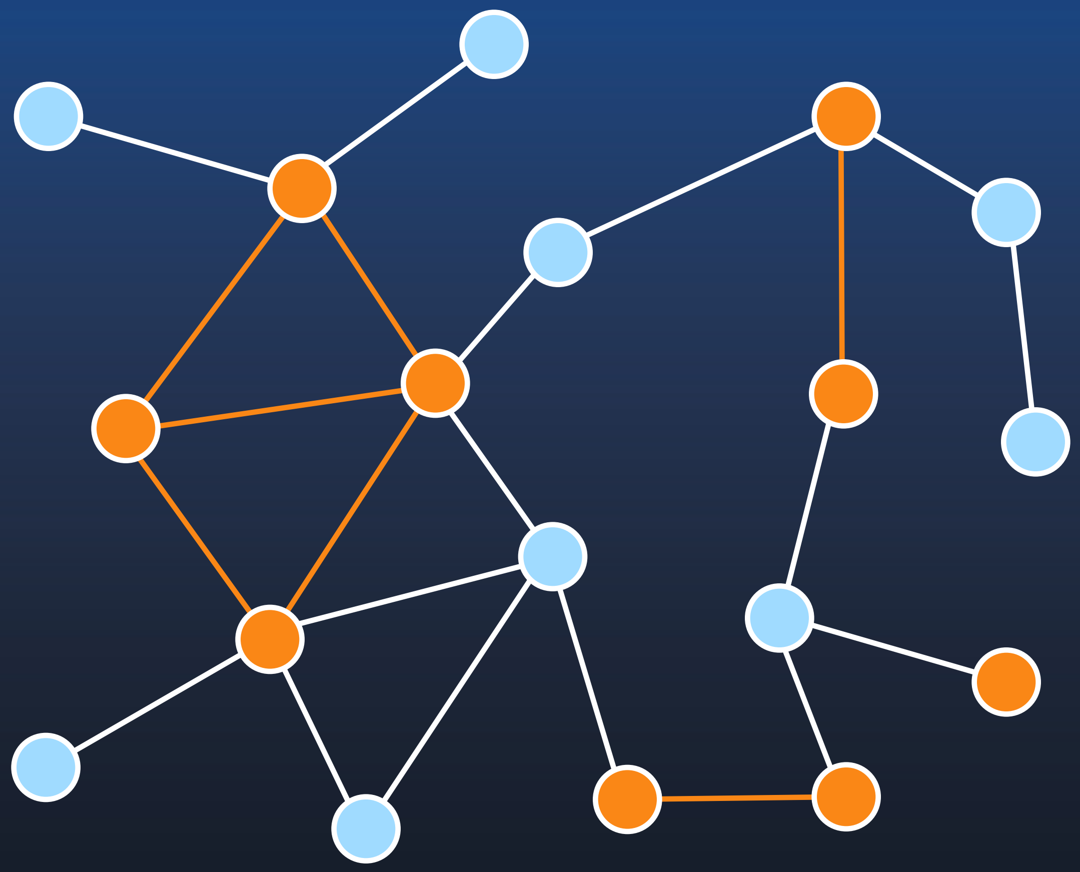


Random Node Sampling

Node-Based Sampling

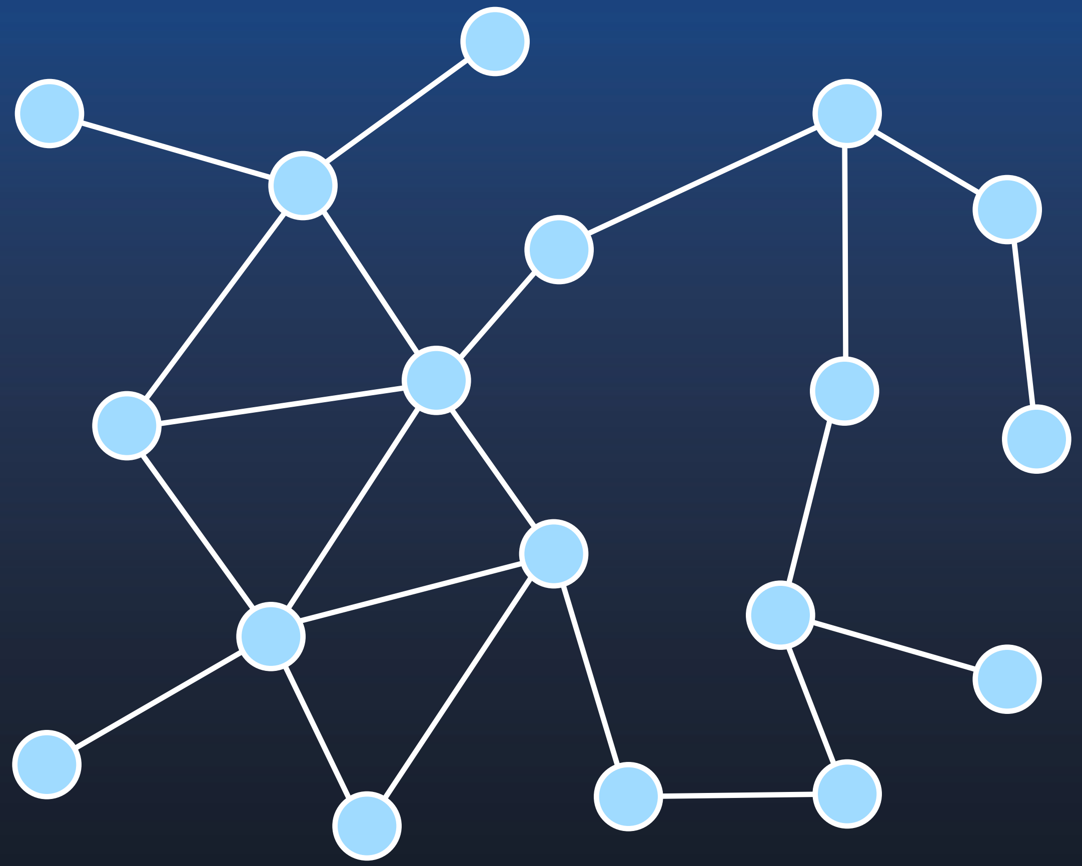


Original Graph

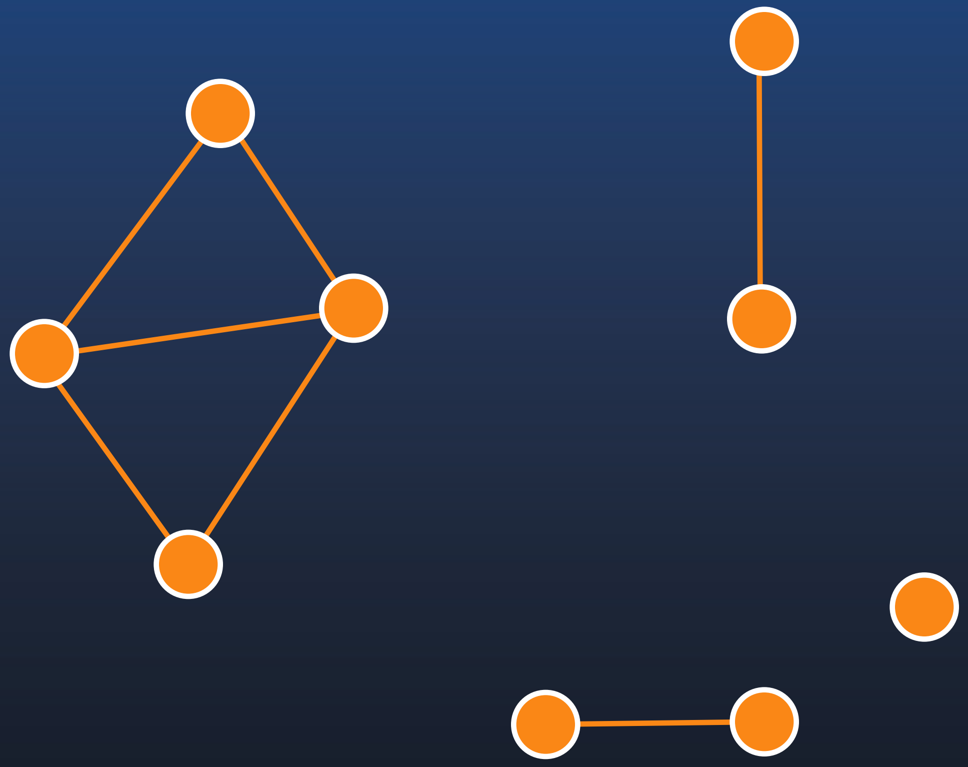


Random Node Sampling

Node-Based Sampling



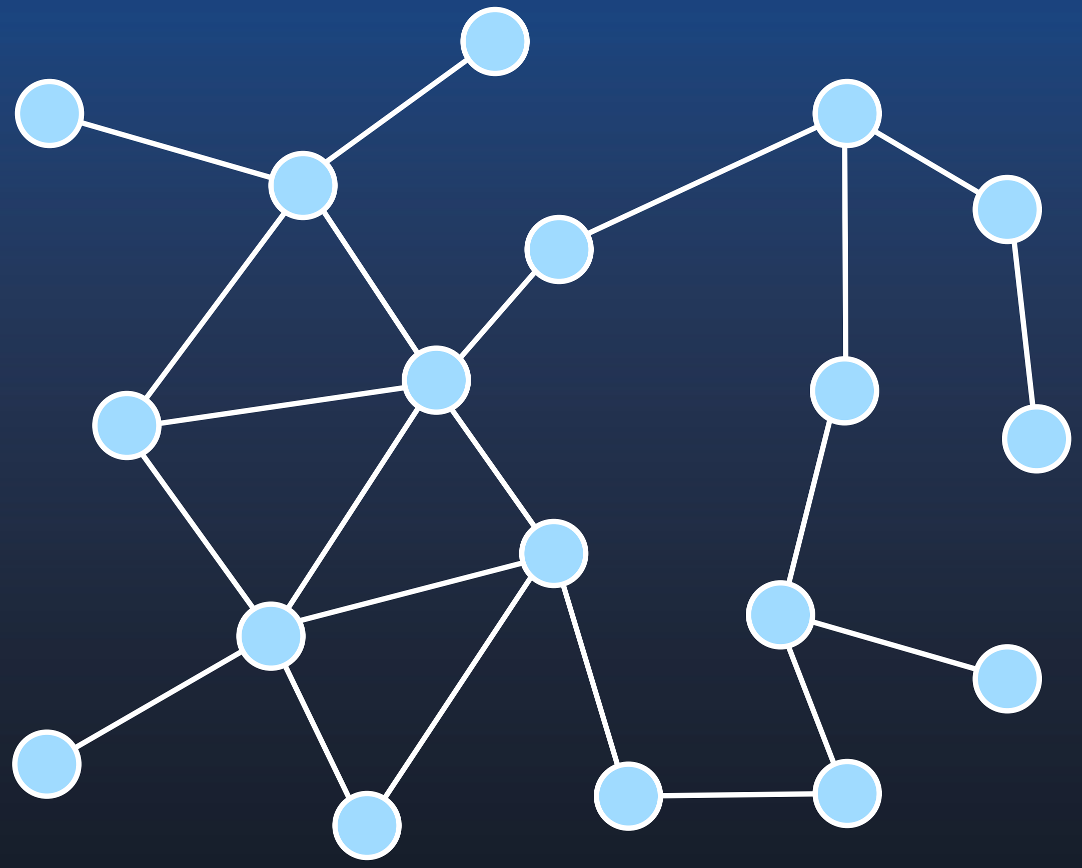
Original Graph



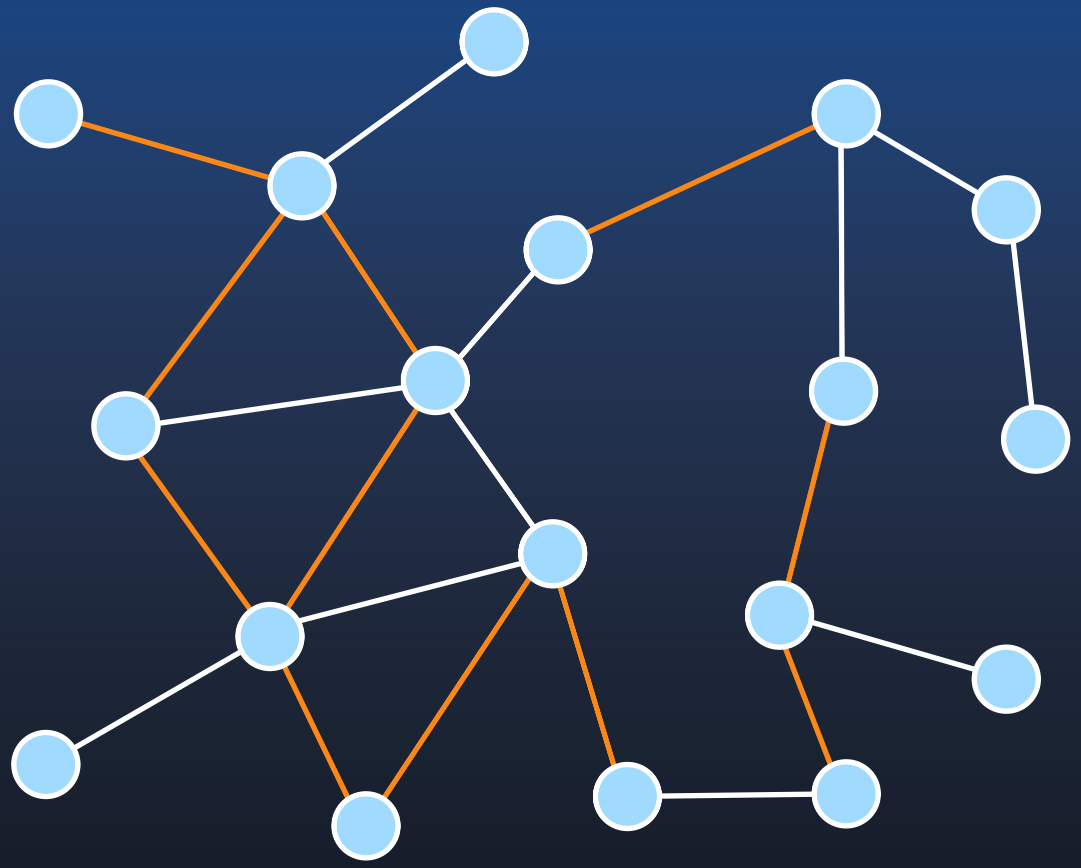
Random Node Sampling

VIS2016

Edge-Based Sampling



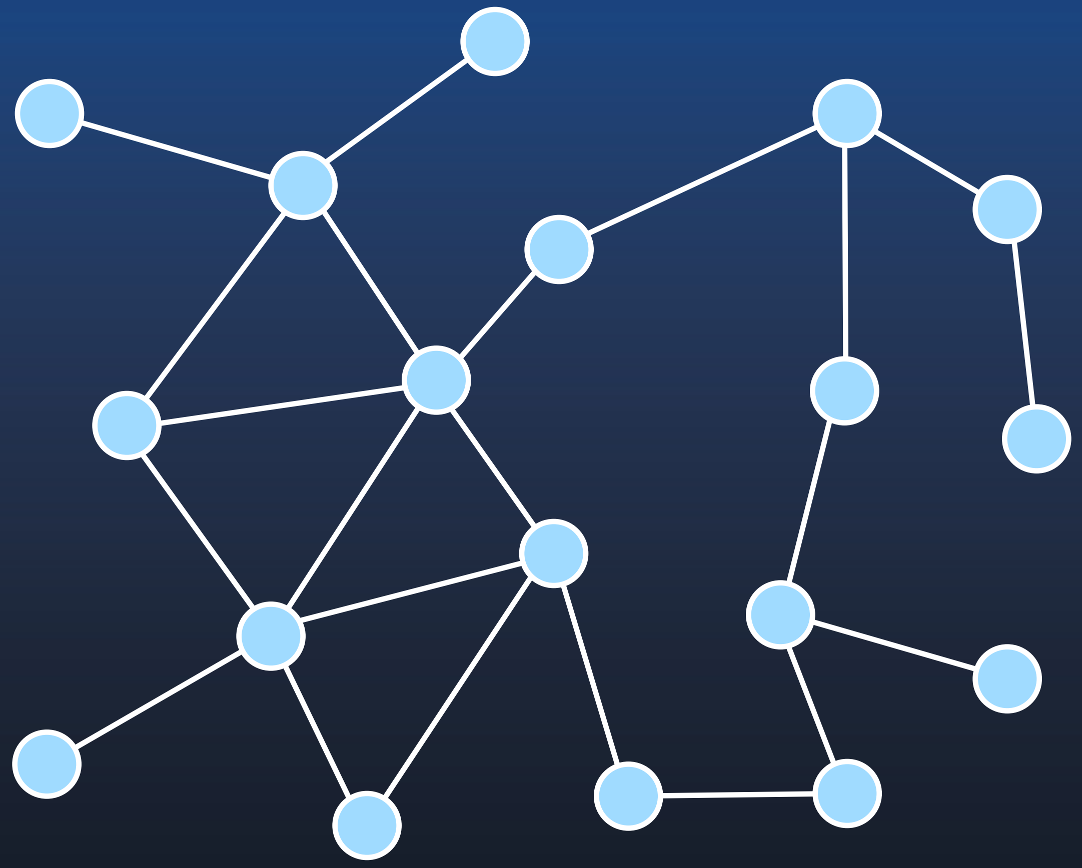
Original Graph



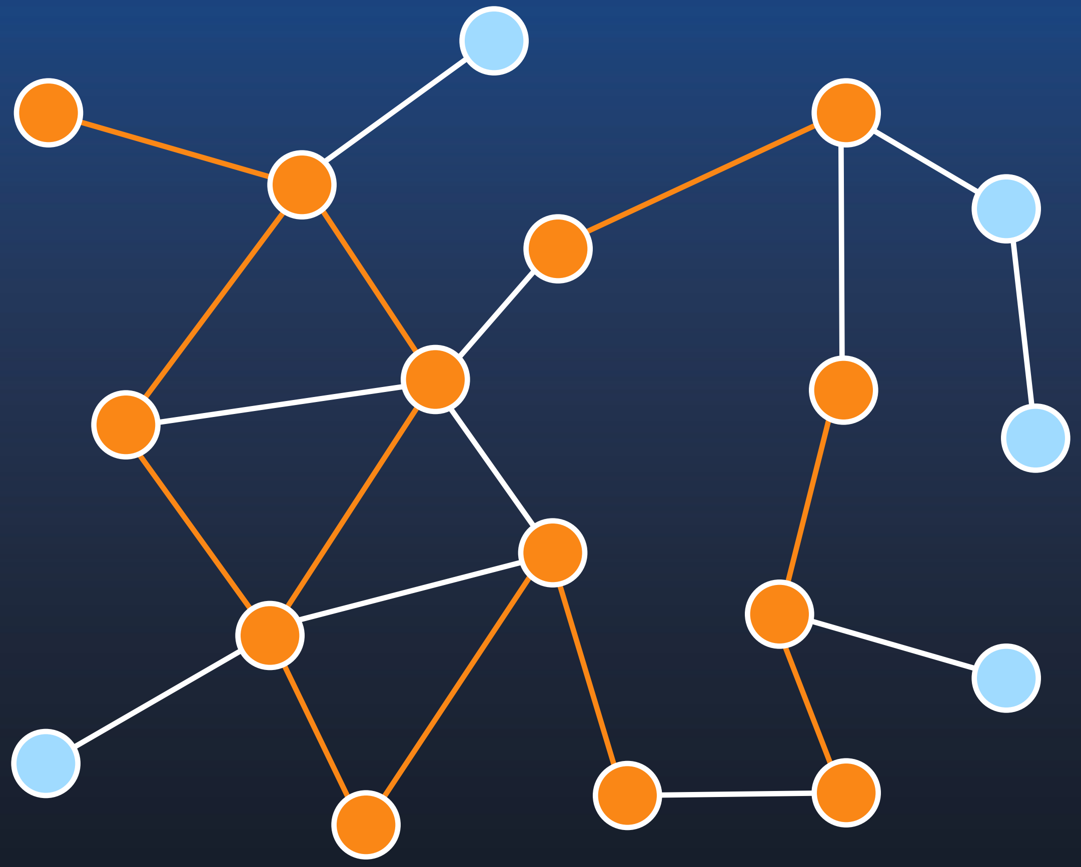
Random Edge Sampling

VIS 2016

Edge-Based Sampling

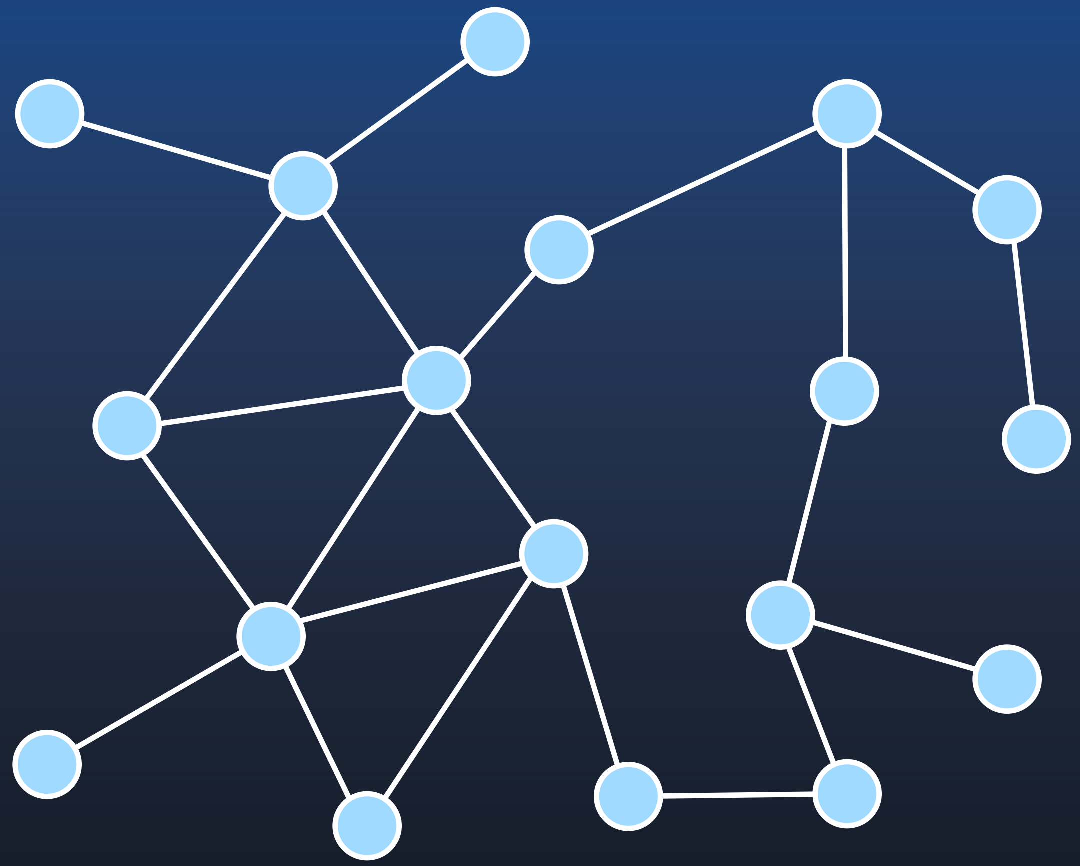


Original Graph

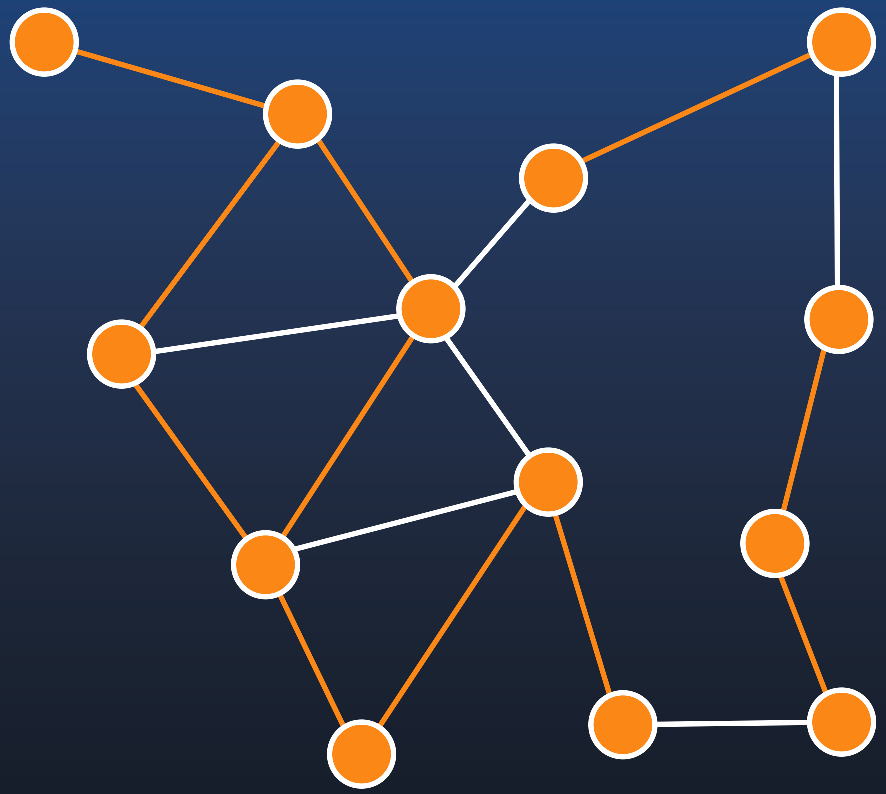


Random Edge Sampling

Edge-Based Sampling

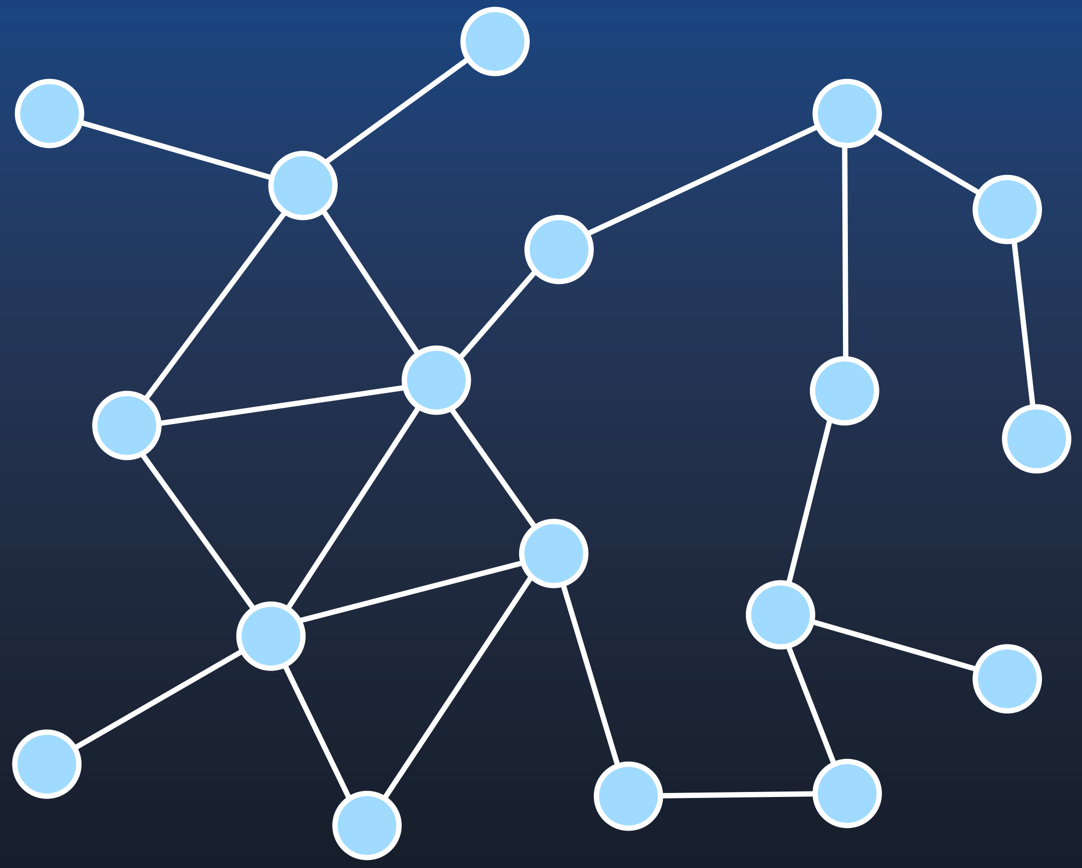


Original Graph

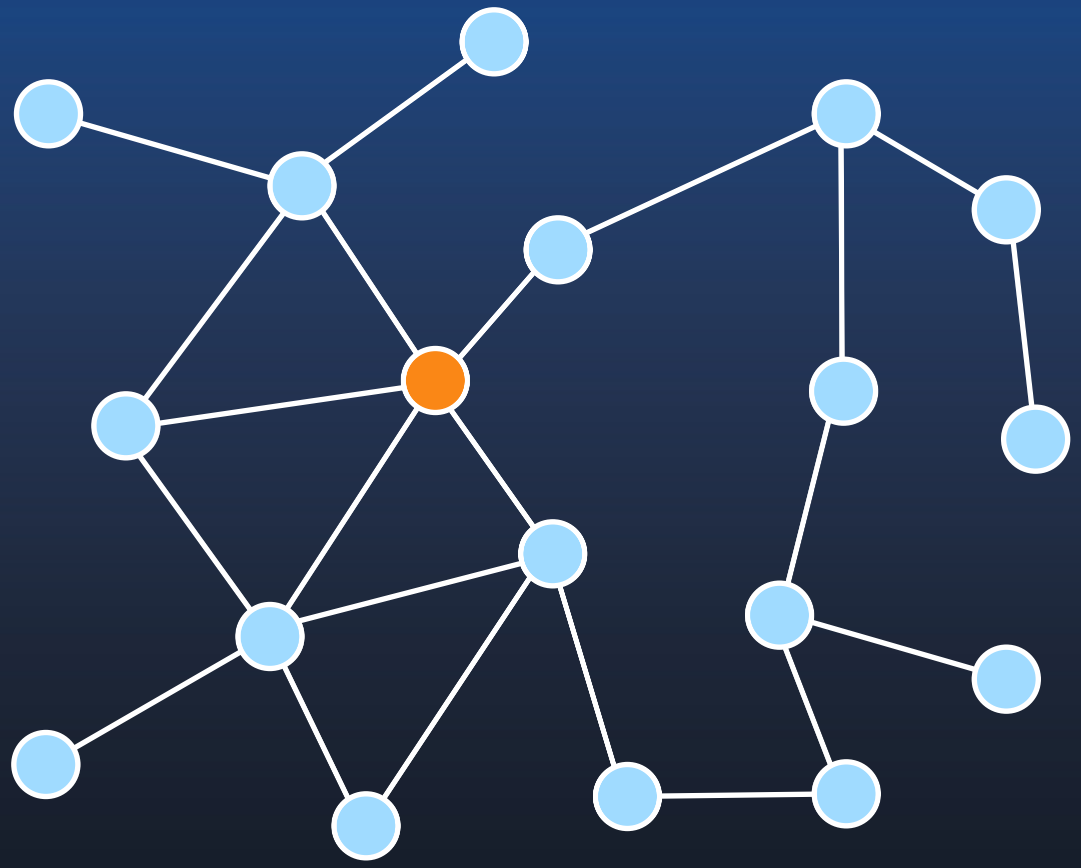


Random Edge Sampling

Traversal-Based Sampling: Random Walk

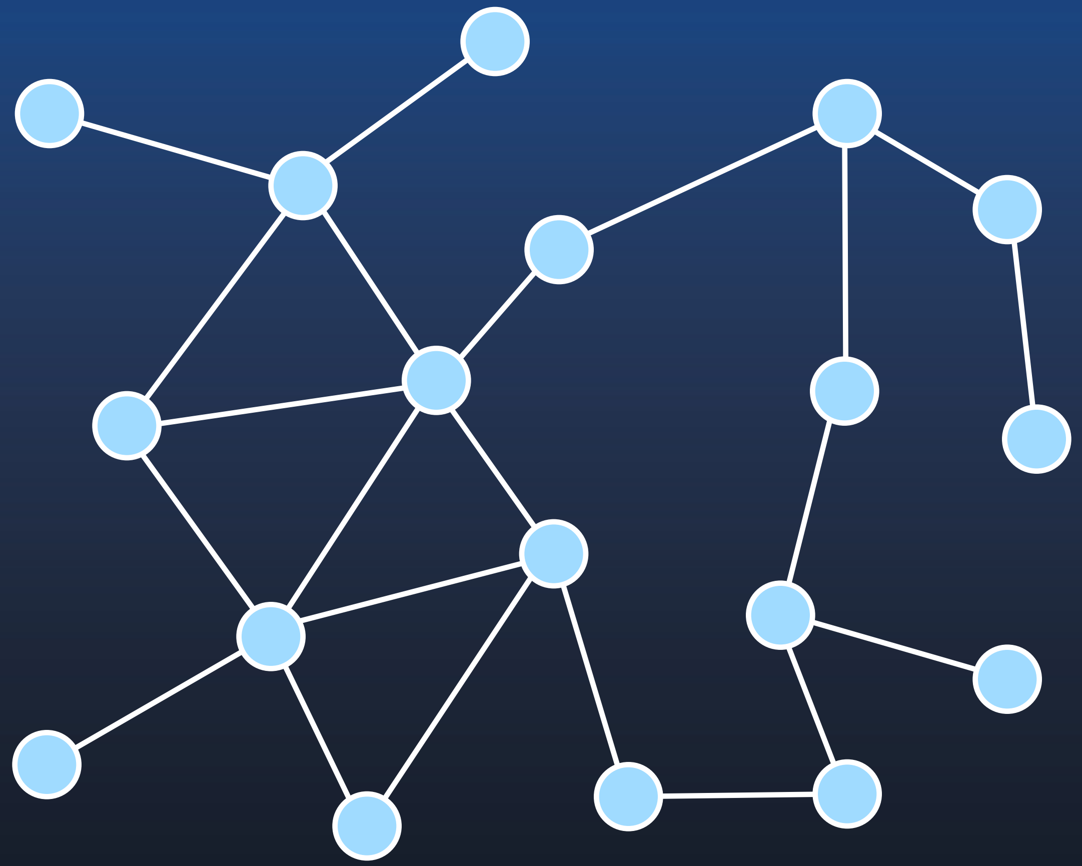


Original Graph

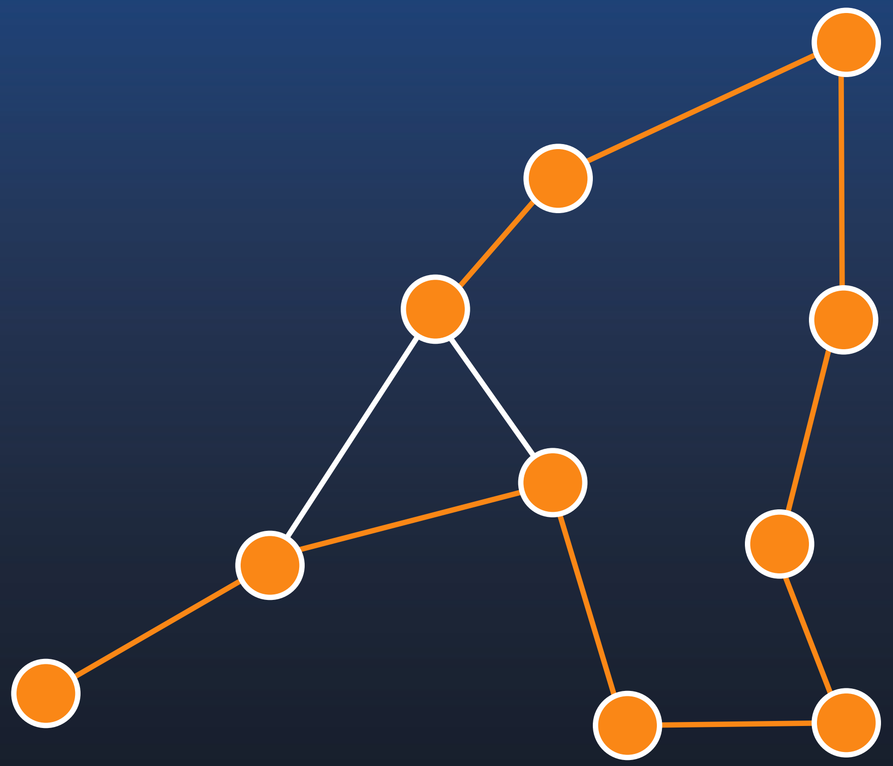


Random Walk

Traversal-Based Sampling: Random Walk

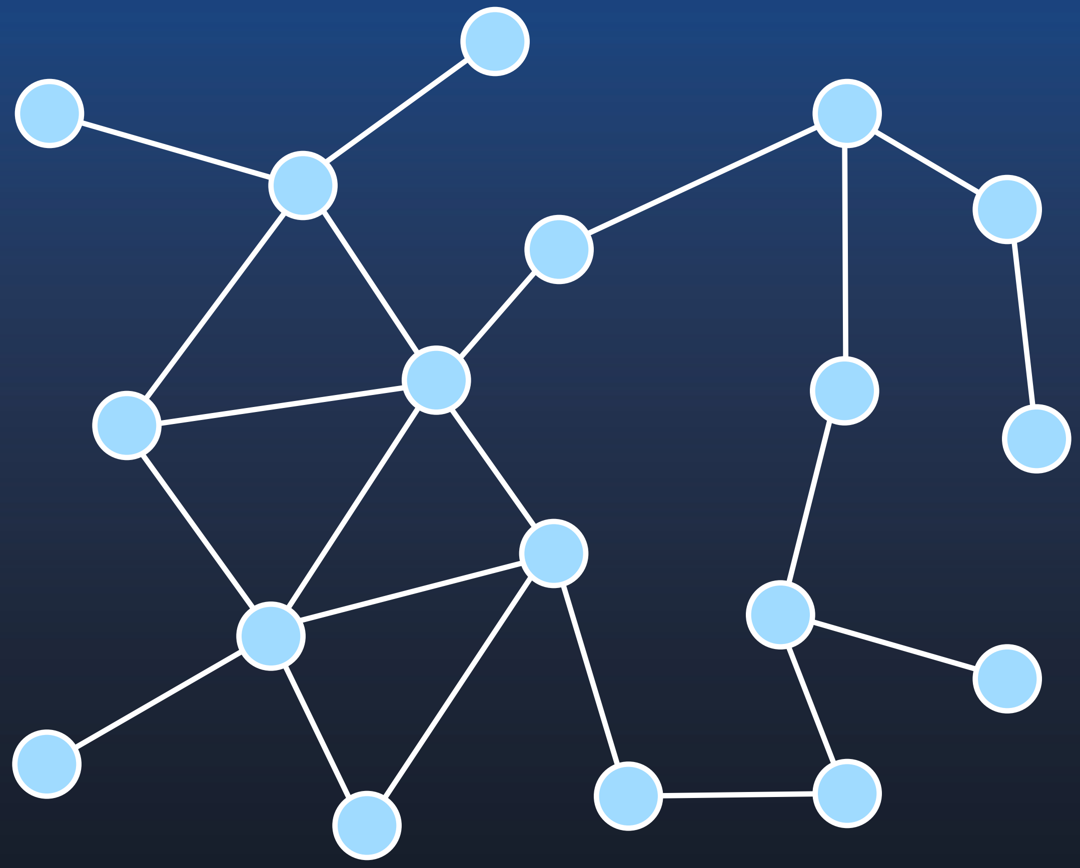


Original Graph

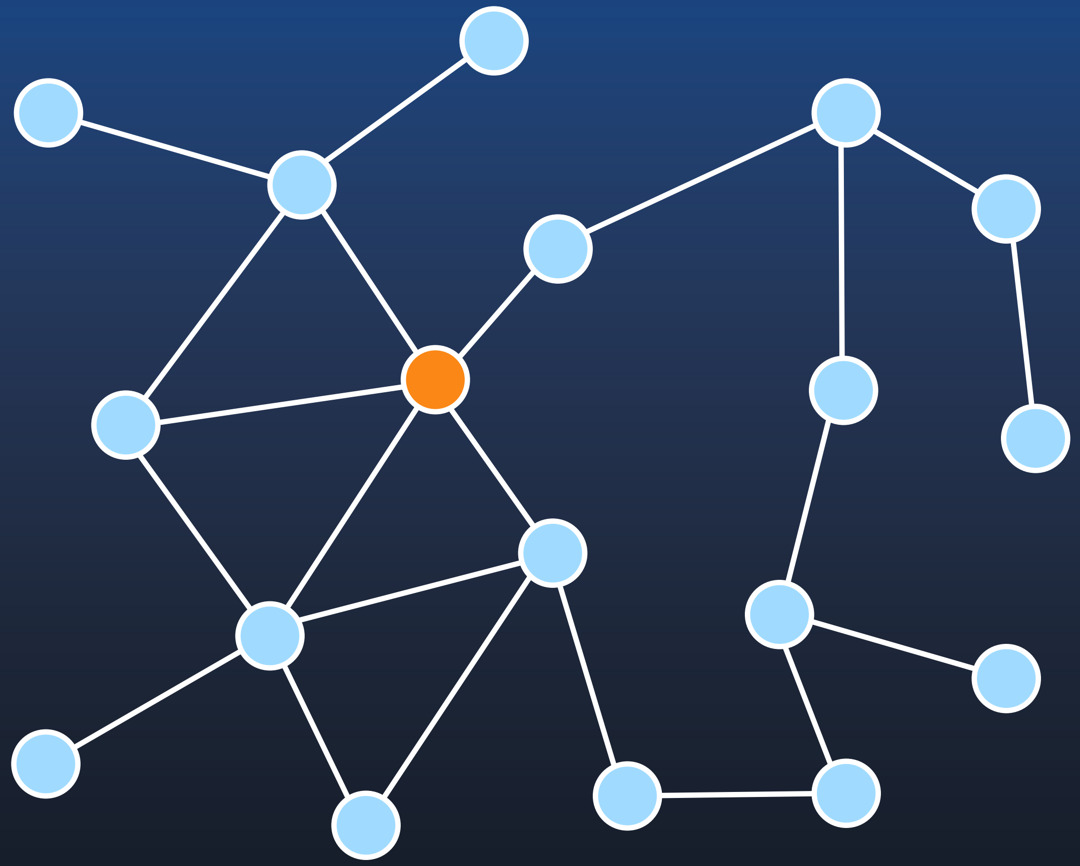


Random Walk

Traversal-Based Sampling: Random Jump

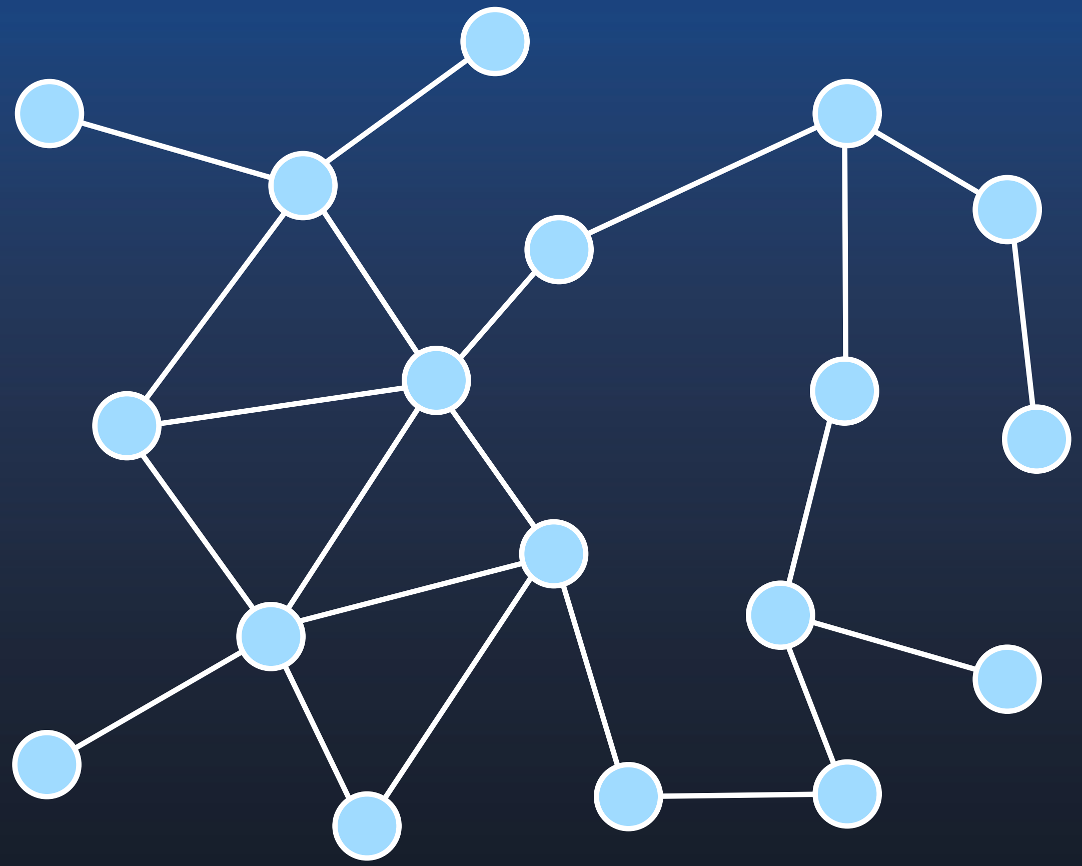


Original Graph

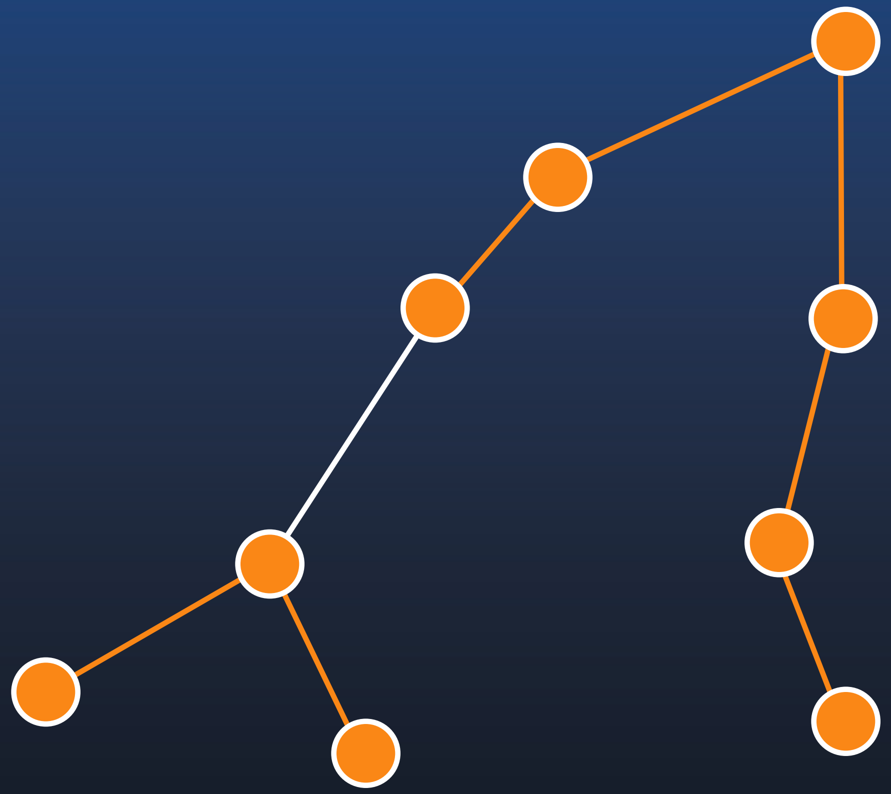


Random Jump

Traversal-Based Sampling: Random Jump

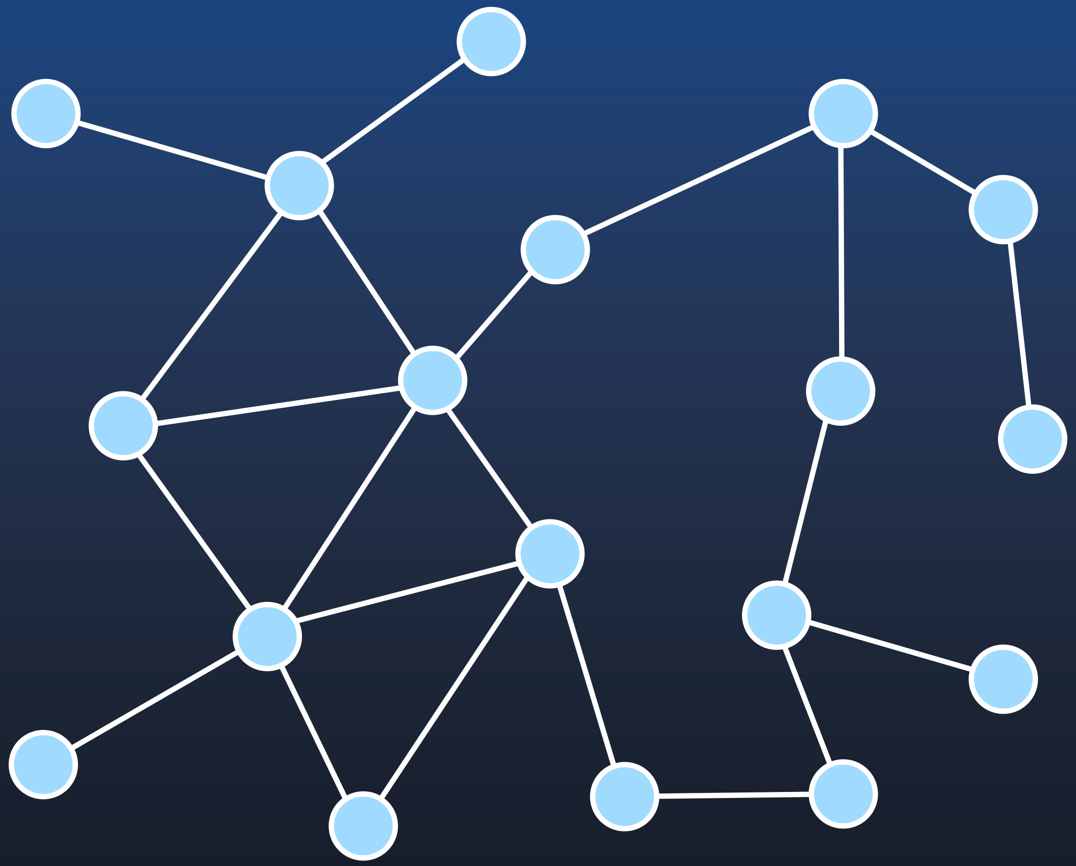


Original Graph

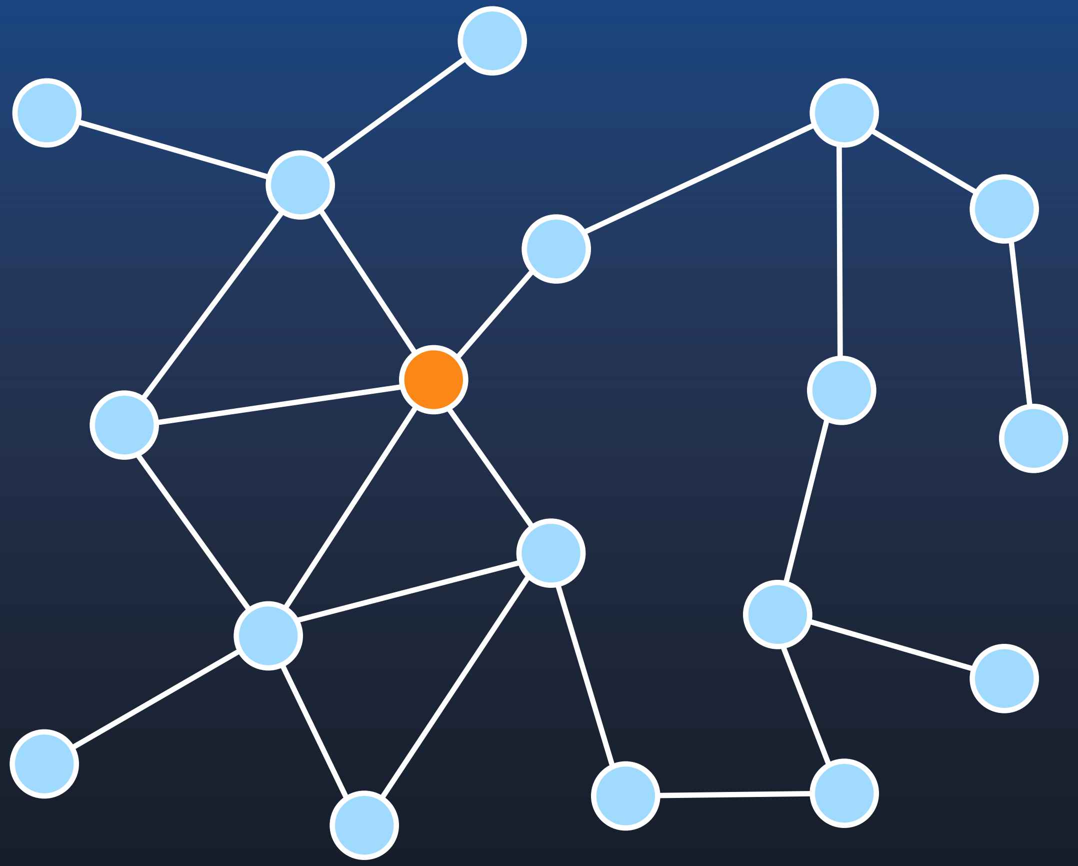


Random Jump

Traversal-Based Sampling: Forest Fire

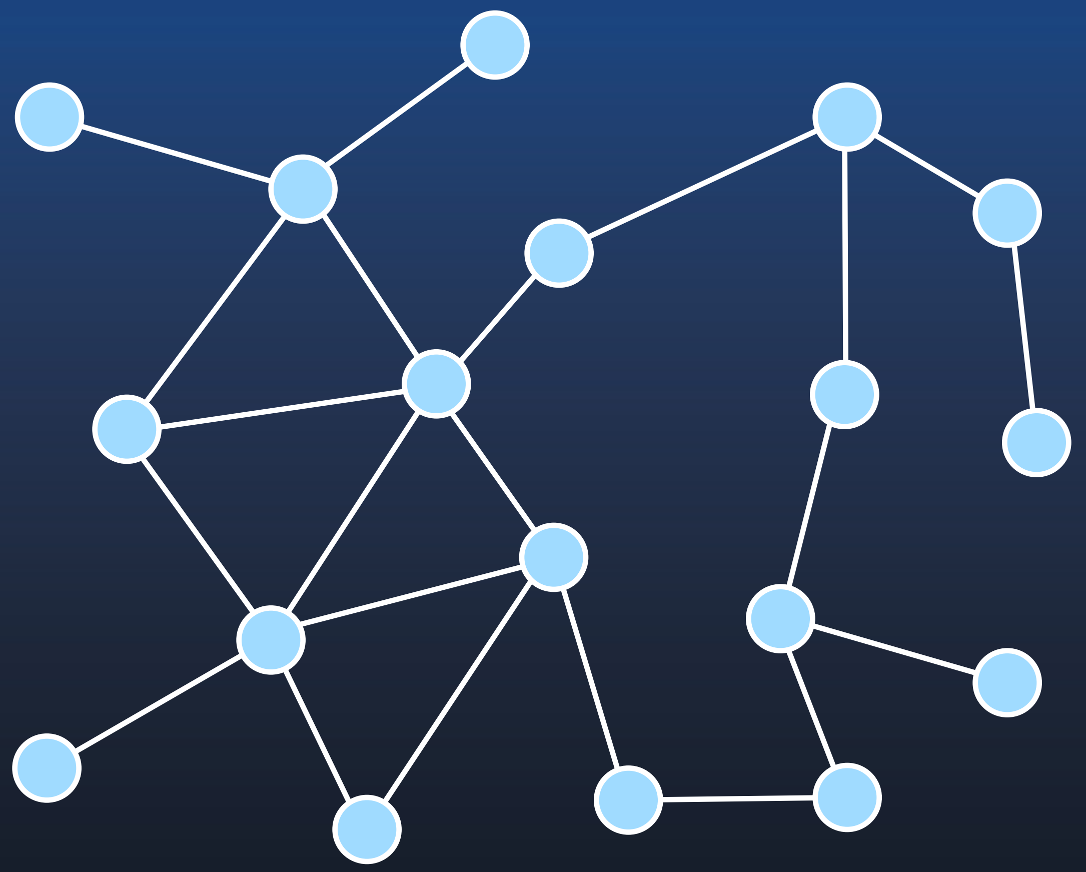


Original Graph

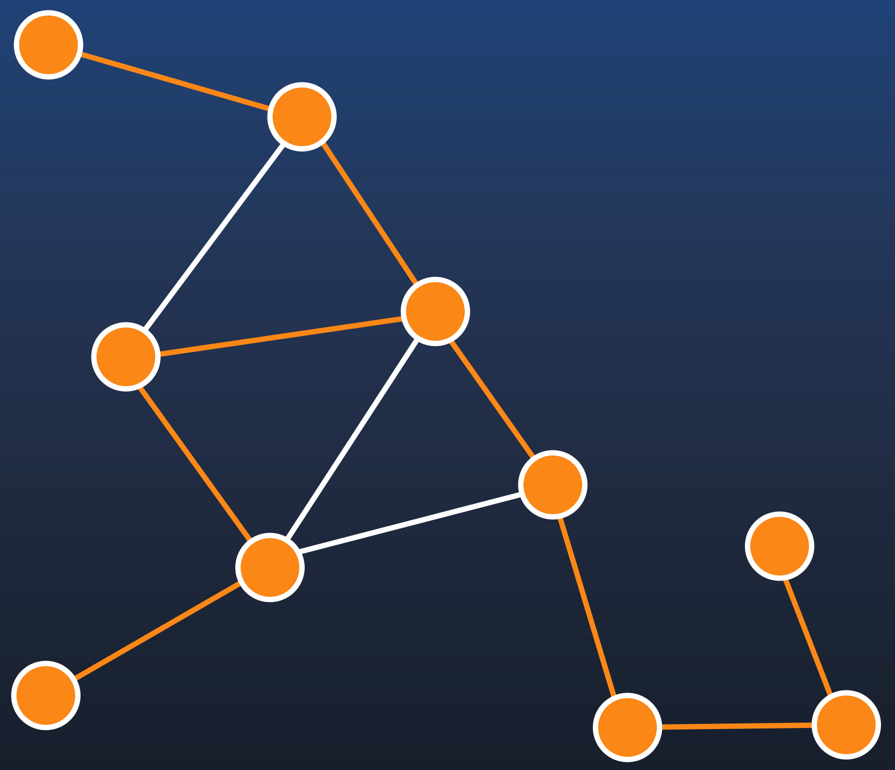


Forest Fire

Traversal-Based Sampling: Forest Fire



Original Graph



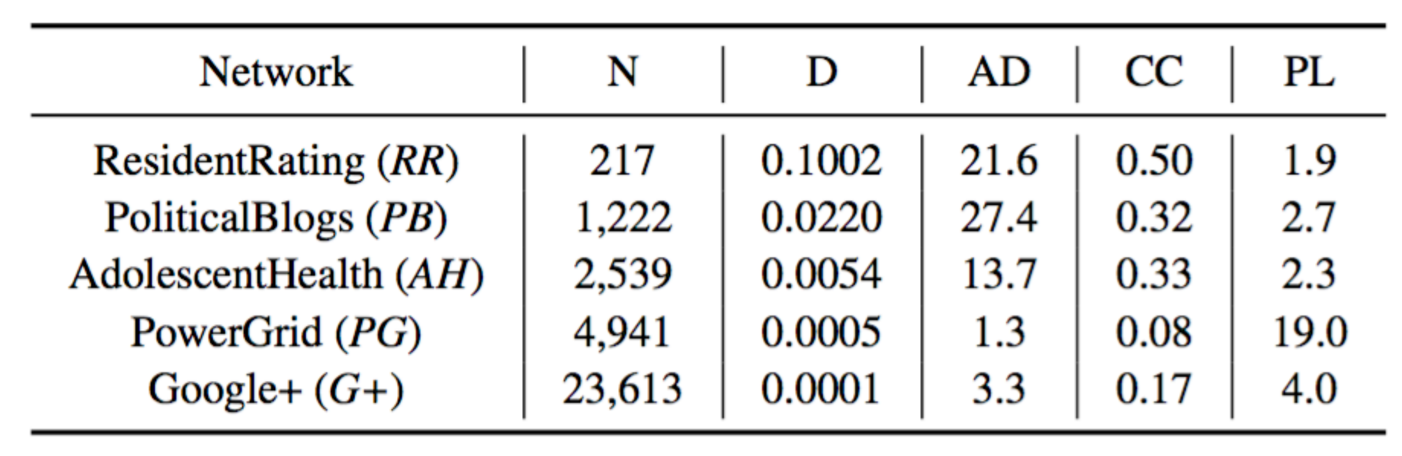
Forest Fire

Outline

- Selected Sampling Methods
- Pilot Study
- Formal Studies
  - Perception of High Degree Nodes
  - Perception of Cluster Quality
  - Perception of Coverage Area

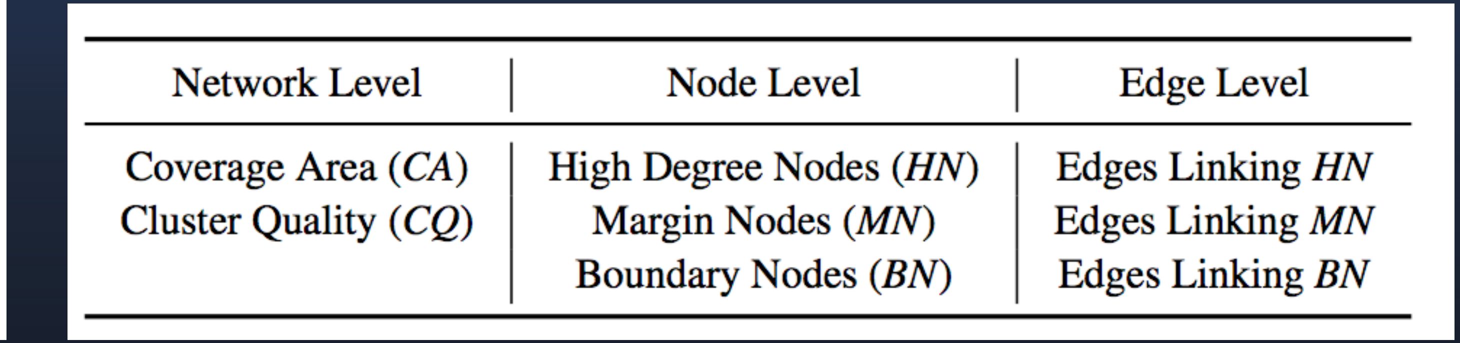
Pilot Study

- Task:
  - Identify the visual factors that strongly influence the representativeness of sampled graphs
  - We also determine the sampling rate used in the formal studies.



Network	N	D	AD	CC	PL
ResidentRating (*RR*)	217	0.1002	21.6	0.50	1.9
PoliticalBlogs (*PB*)	1,222	0.0220	27.4	0.32	2.7
AdolescentHealth (*AH*)	2,539	0.0054	13.7	0.33	2.3
PowerGrid (*PG*)	4,941	0.0005	1.3	0.08	19.0
Google+ (*G+*)	23,613	0.0001	3.3	0.17	4.0

Dataset: 5 Real-World Graphs



Network Level	Node Level	Edge Level
Coverage Area (*CA*)	High Degree Nodes (*HN*)	Edges Linking *HN*
Cluster Quality (*CQ*)	Margin Nodes (*MN*)	Edges Linking *MN*
	Boundary Nodes (*BN*)	Edges Linking *BN*

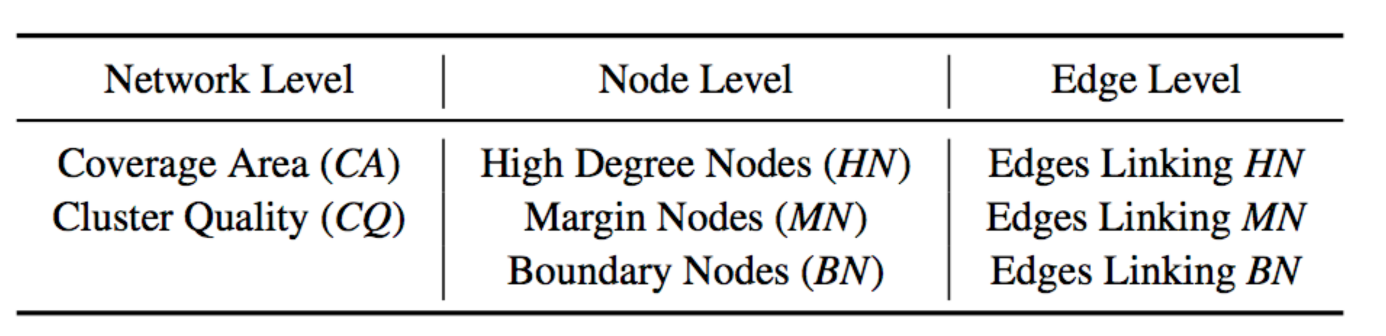
Visual Factor Candidates

Pilot Study

- Task:
  - Identify the visual factors that strongly influence the representativeness of sampled graphs
  - We also determine the sampling rate used in the formal studies.

High Degree Nodes
Cluster Quality
Coverage Area

Results (key visual factors)



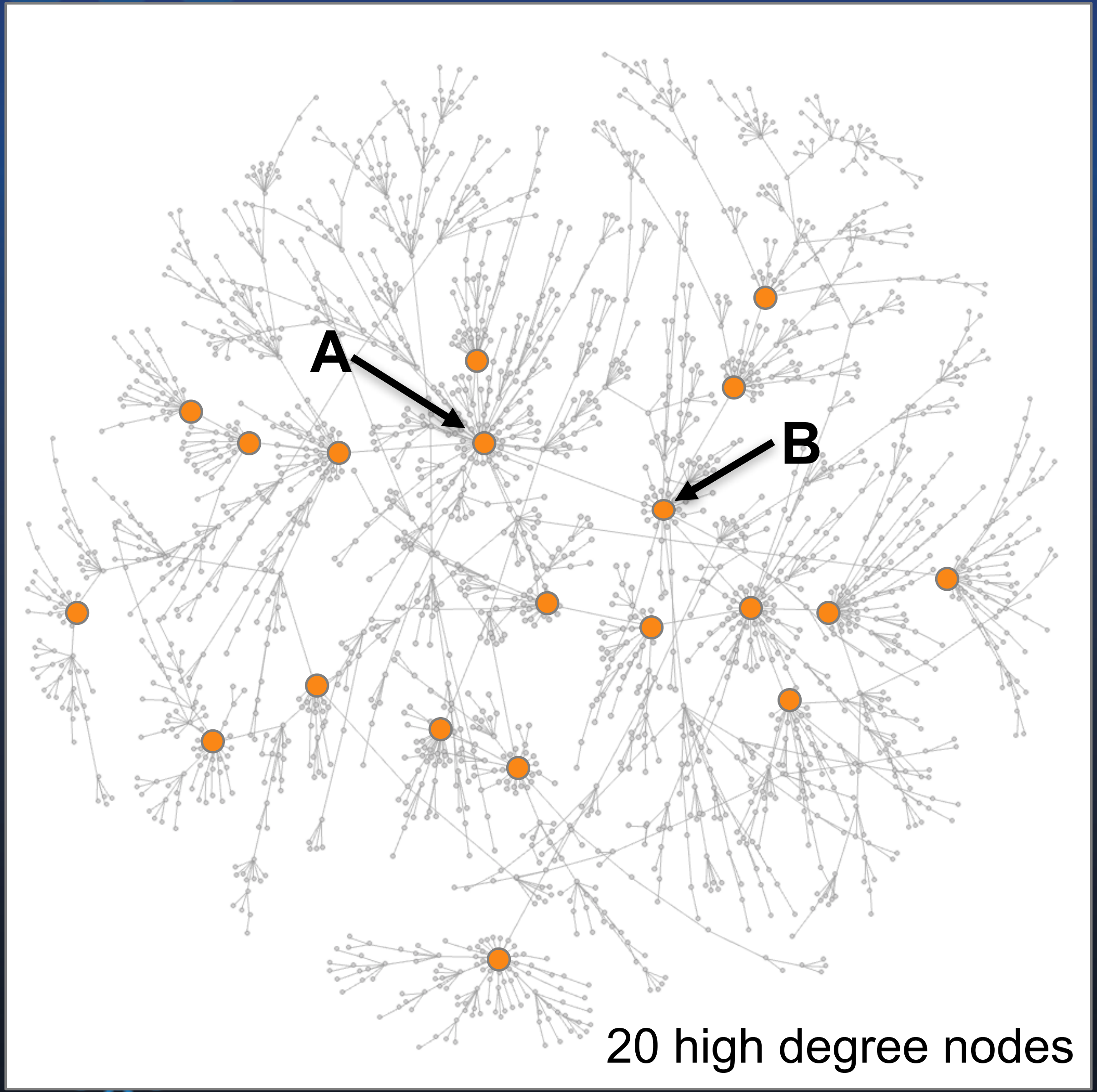
Network Level	Node Level	Edge Level
Coverage Area (*CA*)	High Degree Nodes (*HN*)	Edges Linking *HN*
Cluster Quality (*CQ*)	Margin Nodes (*MN*)	Edges Linking *MN*
	Boundary Nodes (*BN*)	Edges Linking *BN*

Visual Factor Candidates

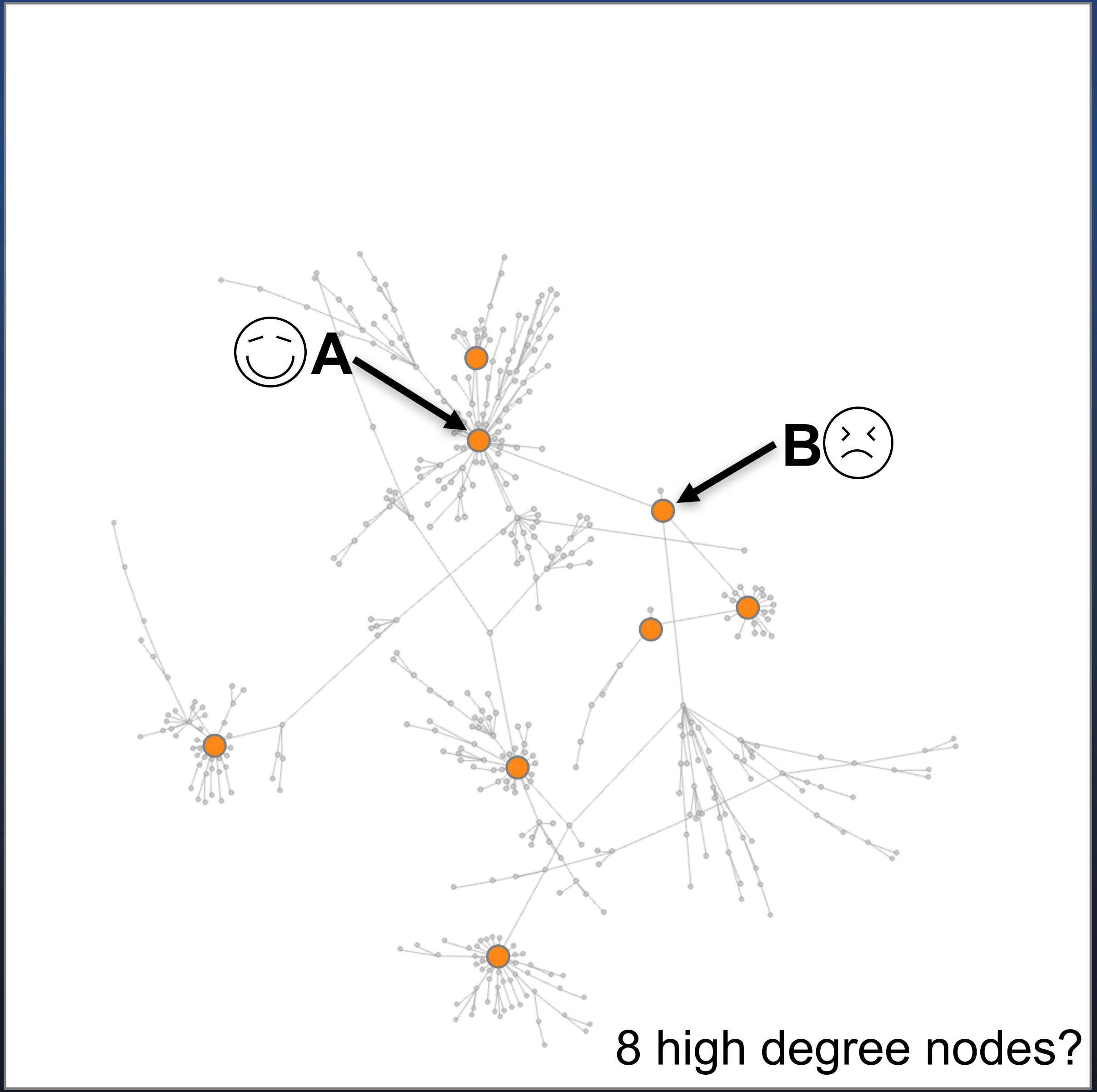
Outline

- Selected Sampling Methods
- Pilot Study
- Formal Studies
  - Perception of High Degree Nodes
  - Perception of Cluster Quality
  - Perception of Coverage Area

Formal Study I: High Degree Nodes

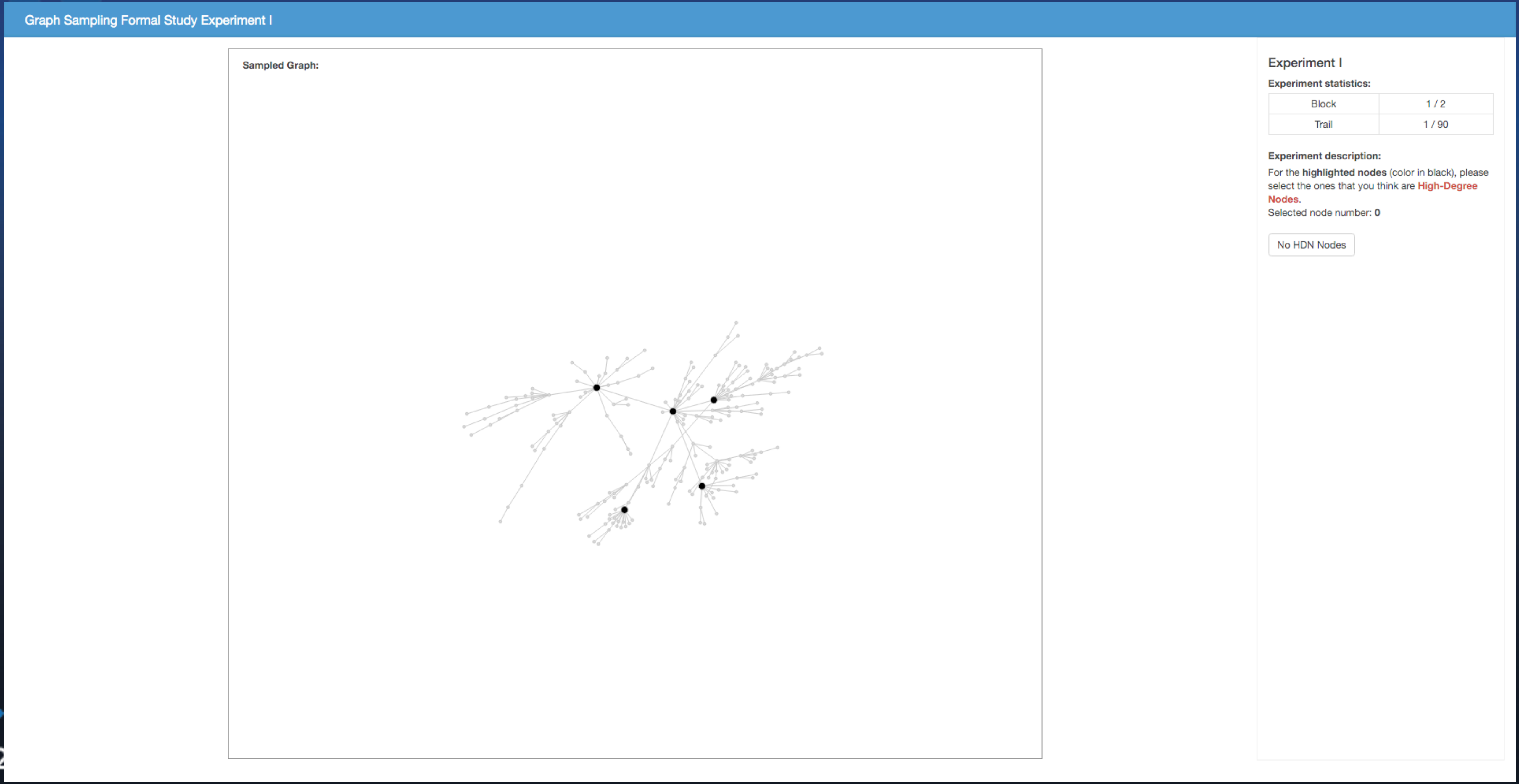


Original Graph



Sampled Graph

Formal Study I: High Degree Nodes



Graph Sampling Formal Study Experiment I

Sampled Graph:

Experiment I

Experiment statistics:

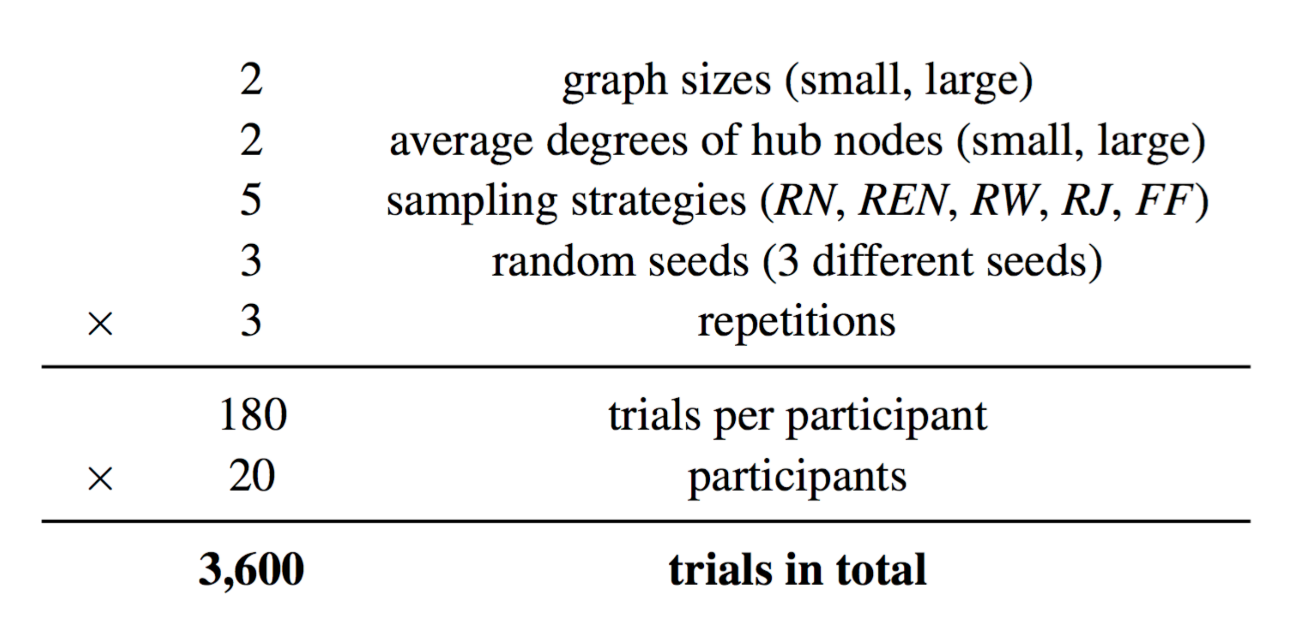
Block	1 / 2
Trail	1 / 90

Experiment description:

For the **highlighted nodes** (color in black), please select the ones that you think are **High-Degree Nodes.**

Selected node number: **0**

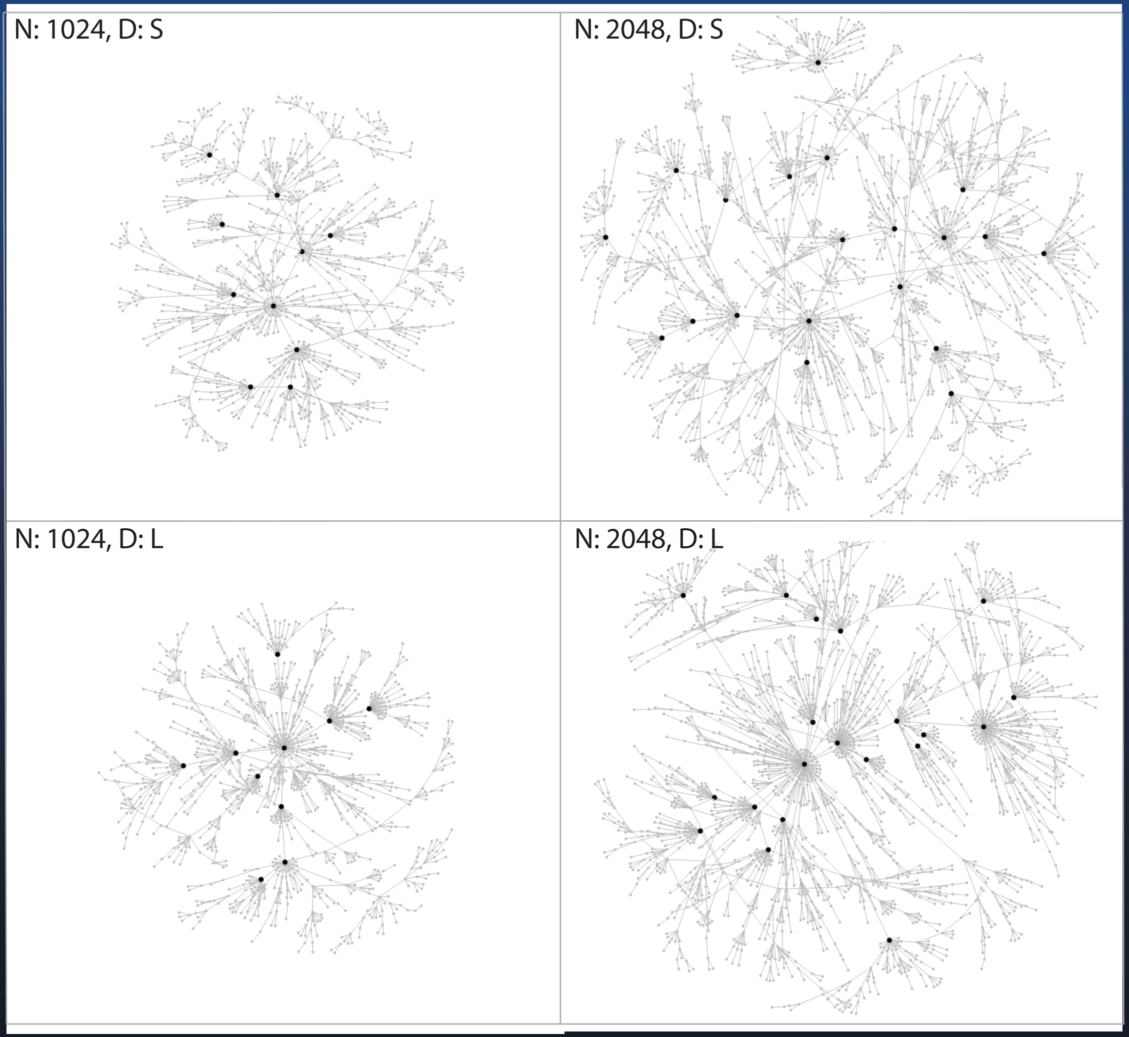
Formal Study I: High Degree Nodes



	2	graph sizes (small, large)
	2	average degrees of hub nodes (small, large)
	5	sampling strategies (*RN*, *REN*, *RW*, *RJ*, *FF*)
	3	random seeds (3 different seeds)
×	3	repetitions
	180	trials per participant
×	20	participants
	3,600	**trials in total**

Experiment Setting

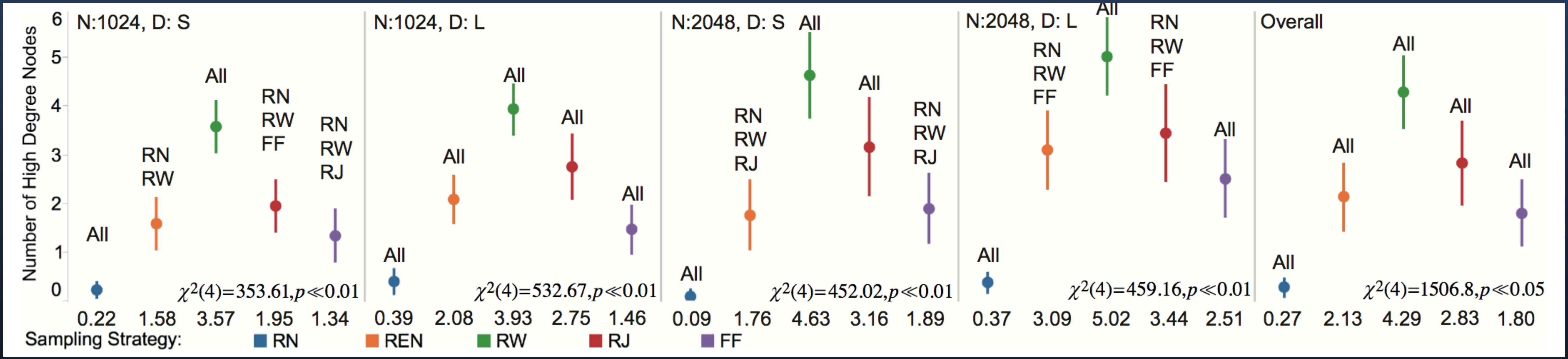
20 high degree nodes



Data Generation

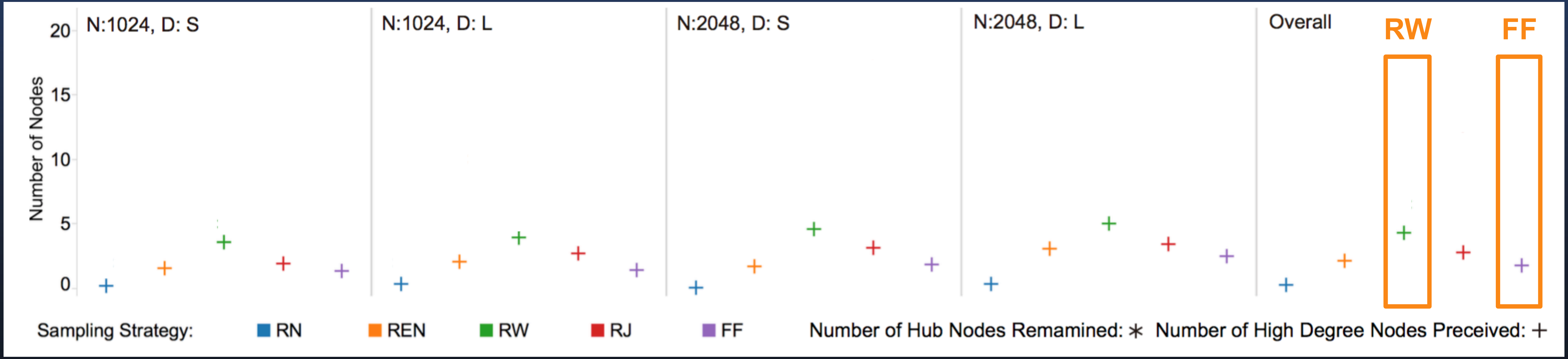
Formal Study I: High Degree Nodes Results

- Discussions:
  - It is easier to perceive high degree nodes in the *RW* Samples
  - It is more difficult to perceive high degree nodes in *RN* Samples
  - Above results hold across datasets



Formal Study I: High Degree Nodes Results

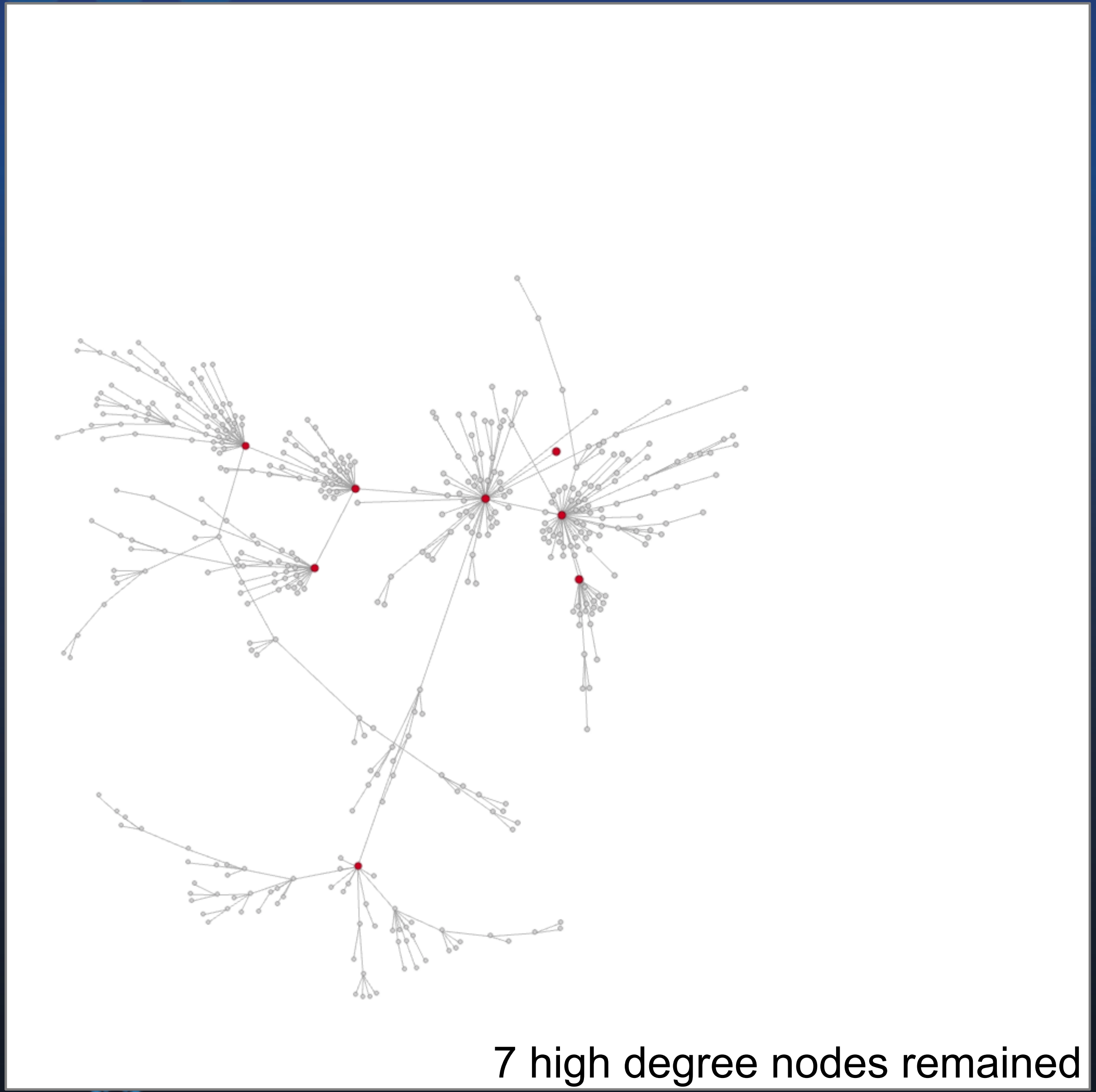
- Discussions:
  - It will be easier to perceive high degree nodes in the *RW* Samples
  - It will be more difficult to perceive high degree nodes in *RN* Samples.
  - Above results hold across datasets



Contradiction with metric-based results!

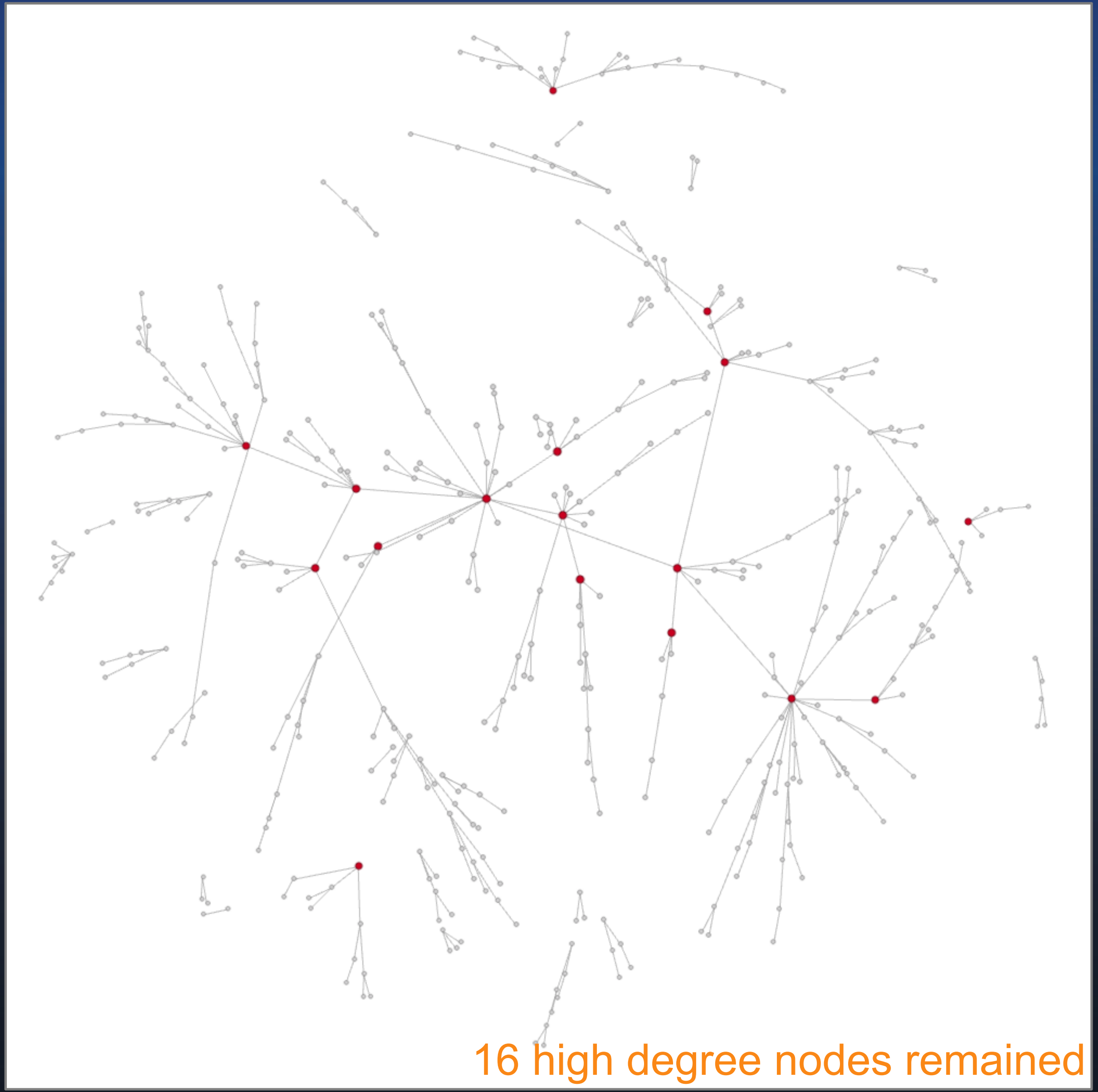
Number of high degree nodes perceived (Visualization): +
Number of high degree nodes remained (Data Mining): *

Formal Study I: High Degree Nodes Results



7 high degree nodes remained

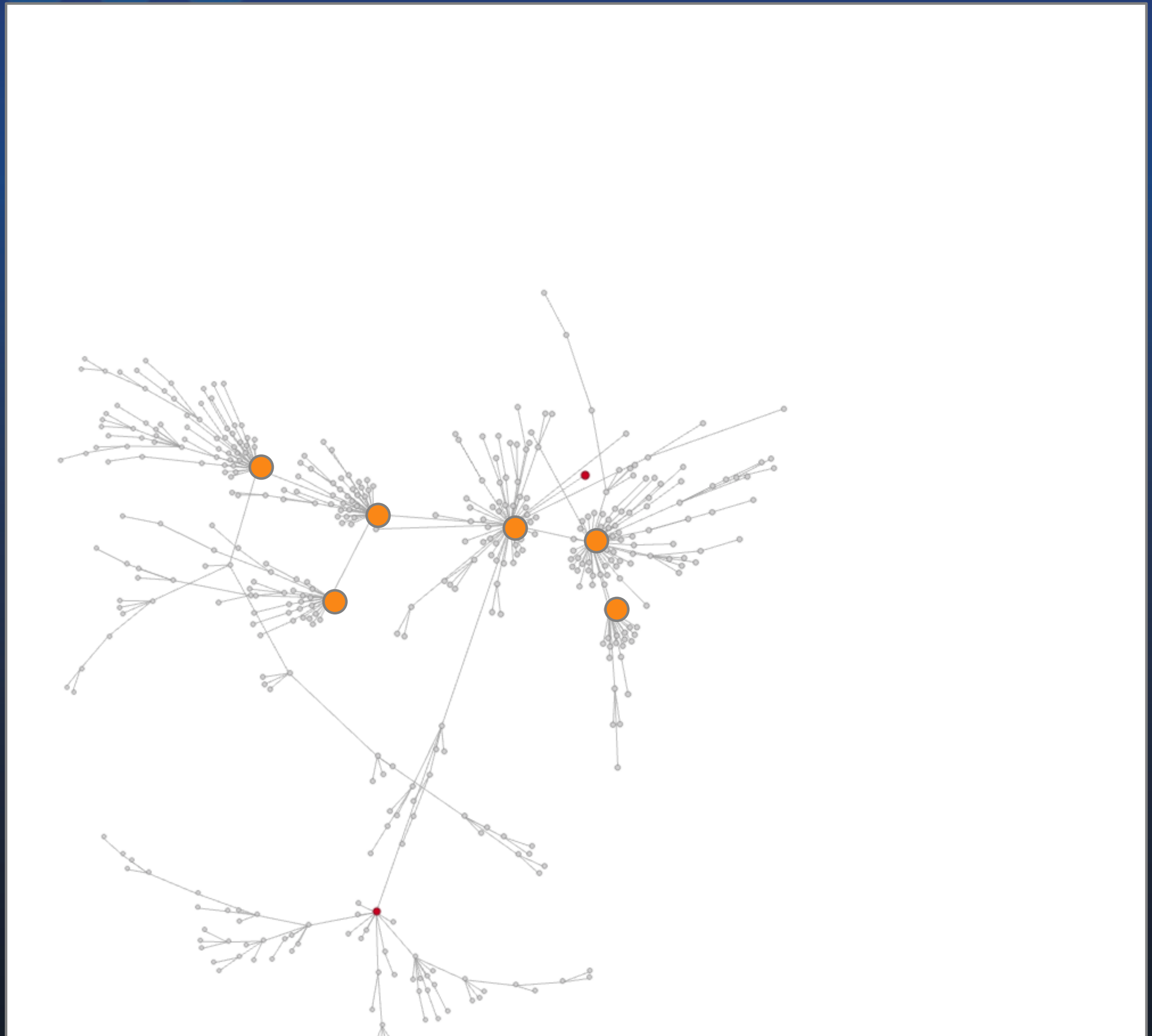
Random Walk (RW)



16 high degree nodes remained

Forest Fire (FF)

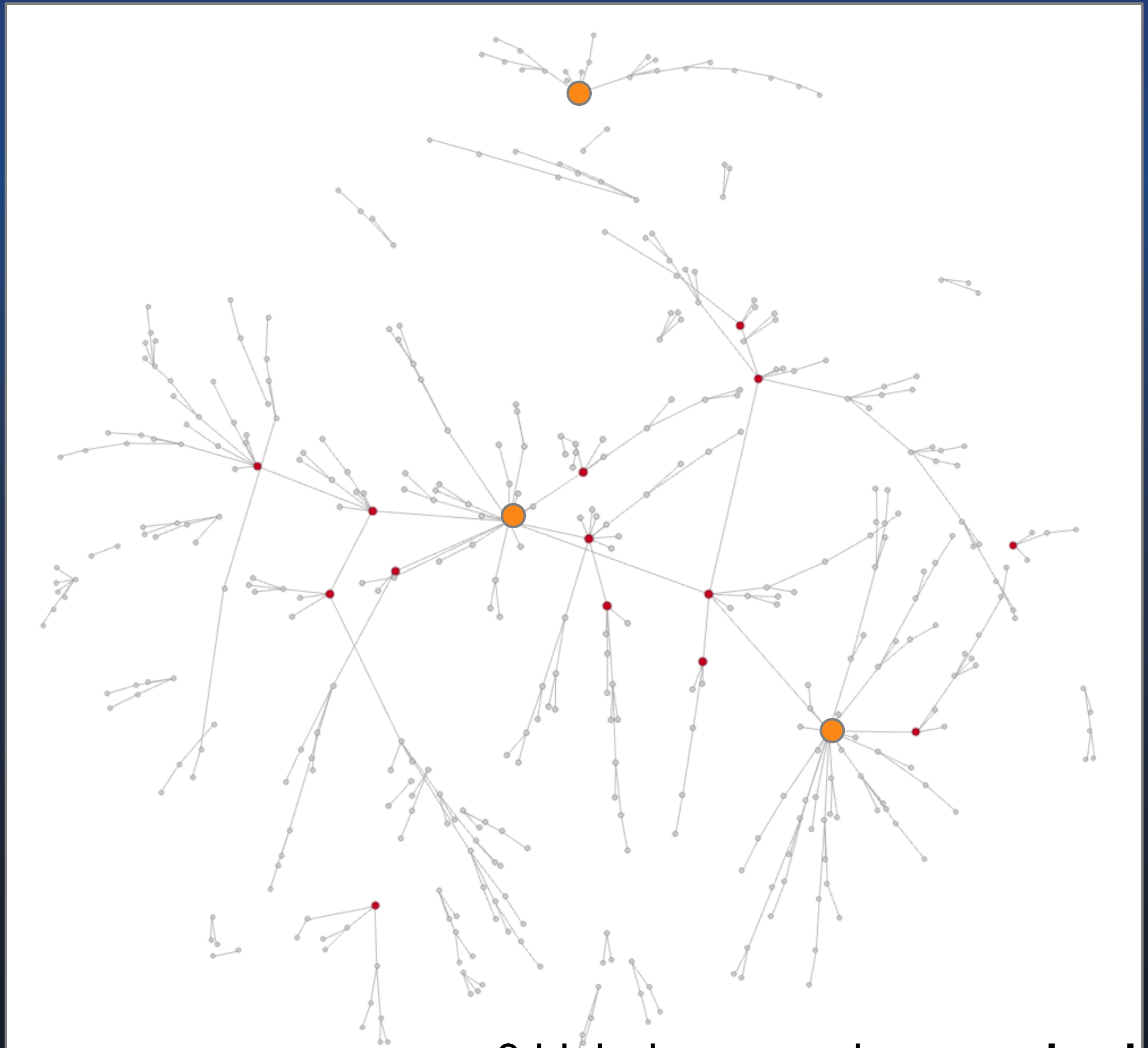
Formal Study I: High Degree Nodes Results



6 high degree nodes **perceived**

~~7 high degree nodes remained~~

Random Walk (RW)



3 high degree nodes **perceived**

16 high degree nodes remained

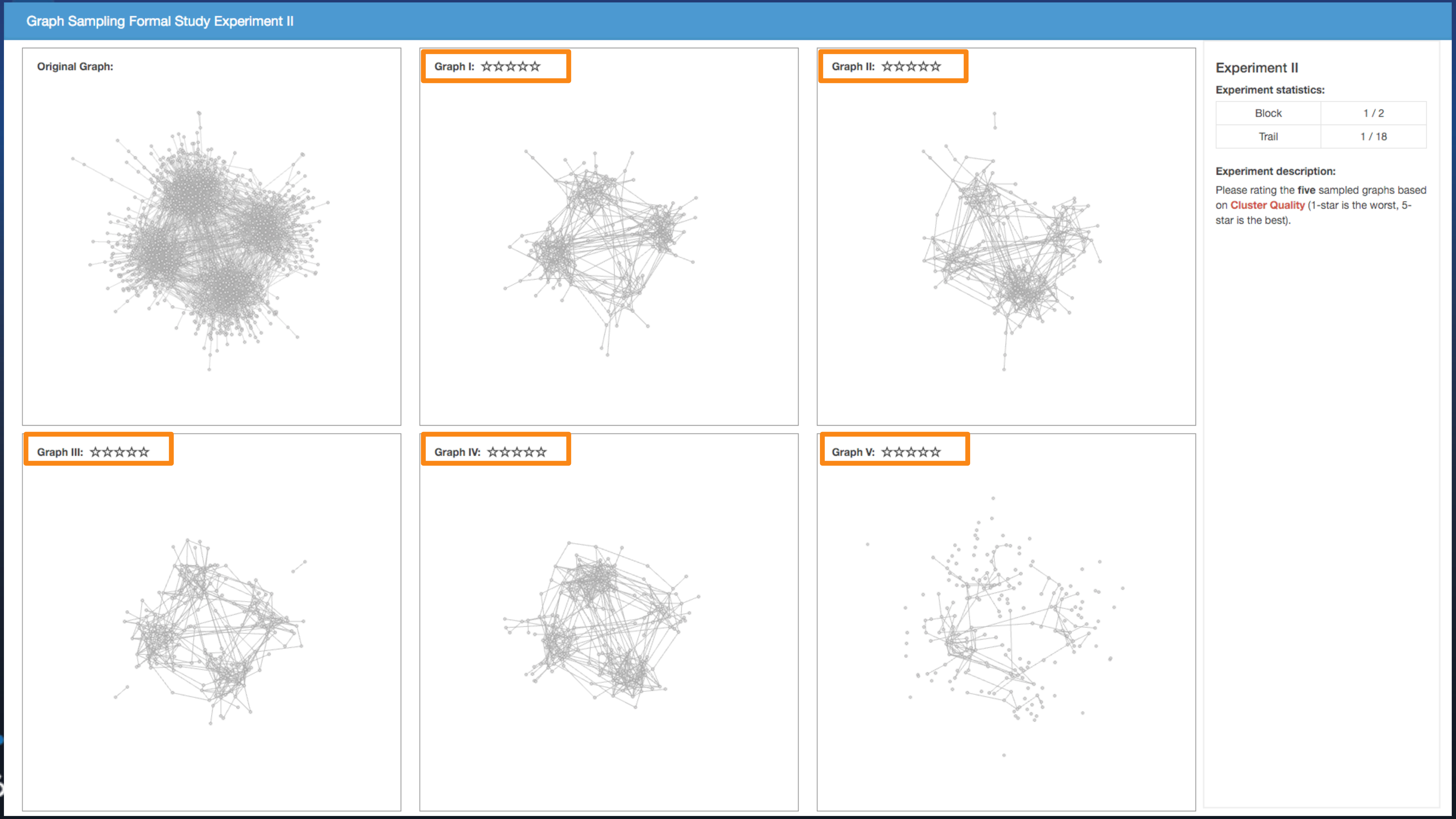
Forest Fire (FF)

VIS2016

Outline

- Selected Sampling Methods
- Pilot Study
- Formal Studies
    - Perception of High Degree Nodes (more high degree nodes are perceived in *RW*)
    - Perception of Cluster Quality
    - Perception of Coverage Area

Formal Study II: Cluster Quality



Graph Sampling Formal Study Experiment II

Original Graph:

Graph I: ☆☆☆☆☆

Graph II: ☆☆☆☆☆

Graph III: ☆☆☆☆☆

Graph IV: ☆☆☆☆☆

Graph V: ☆☆☆☆☆

Experiment II

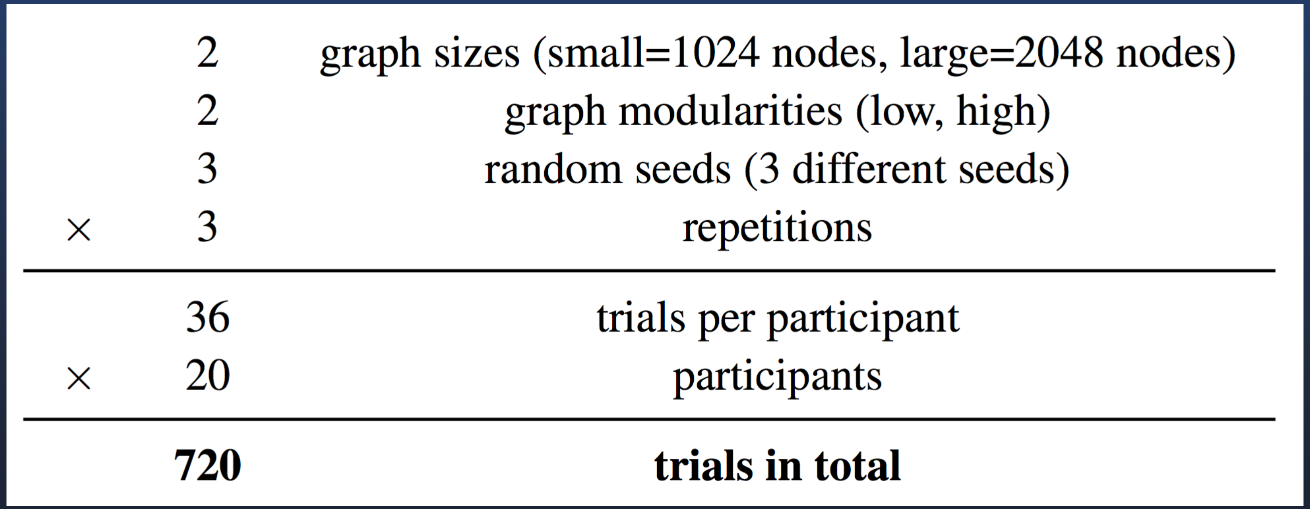
Experiment statistics:

Block	1 / 2
Trail	1 / 18

Experiment description:

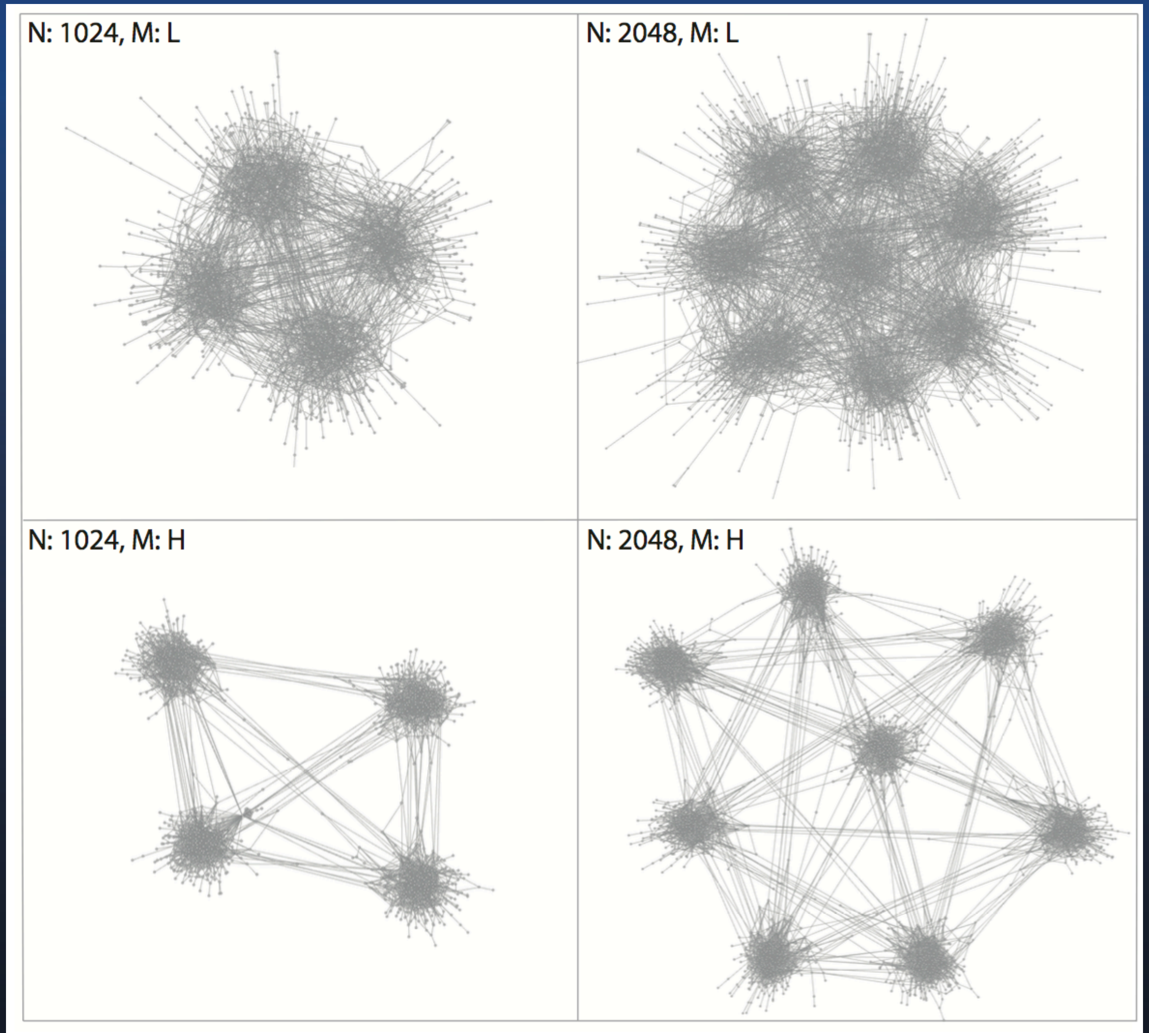
Please rating the **five** sampled graphs based on **Cluster Quality** (1-star is the worst, 5-star is the best).

Formal Study II: Cluster Quality



	2	graph sizes (small=1024 nodes, large=2048 nodes)
	2	graph modularities (low, high)
	3	random seeds (3 different seeds)
×	3	repetitions
	36	trials per participant
×	20	participants
	720	**trials in total**

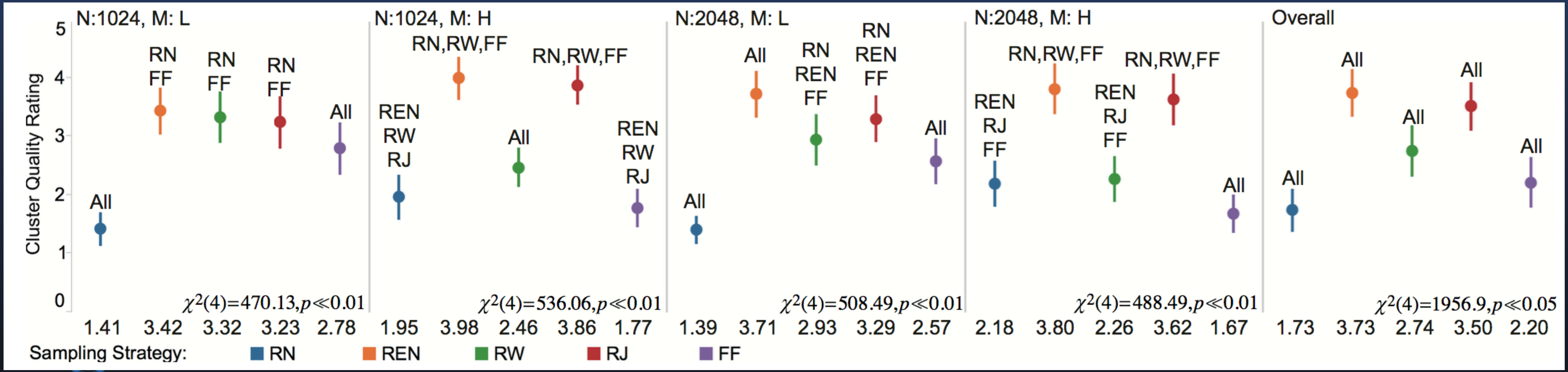
Experiment Setting



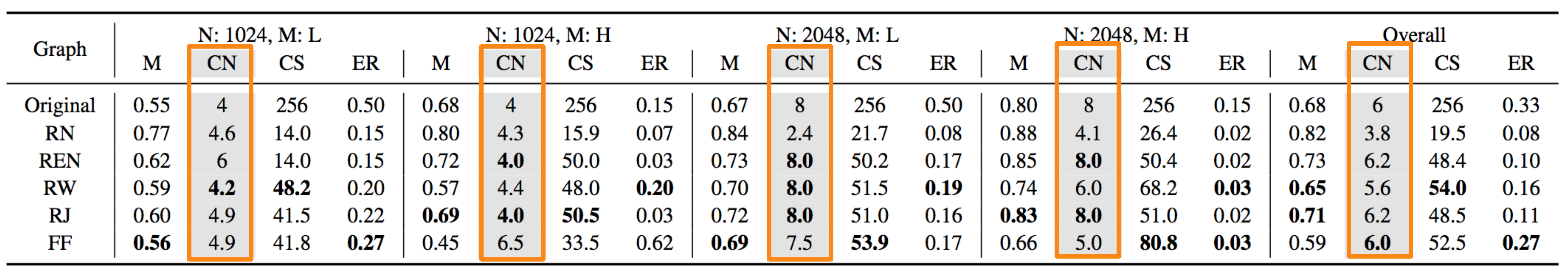
Data Generation

Formal Study II: Cluster Quality Results

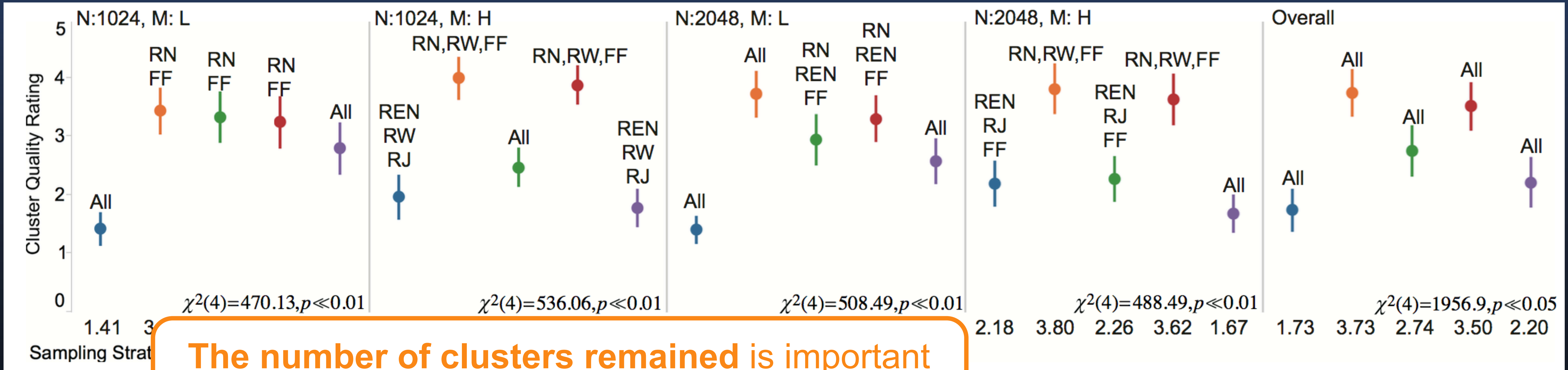
- Discussions:
  - *RE* and *RJ* best preserve the perceived cluster quality in samples
  - *RN and FF* struggles in preserving the perceived cluster quality
  - The performance of *RW* and *FF* depends on graph modularity



Formal Study II: Cluster Quality Results



Graph	N: 1024, M: L				N: 1024, M: H				N: 2048, M: L				N: 2048, M: H				Overall			
	M	CN	CS	ER	M	CN	CS	ER	M	CN	CS	ER	M	CN	CS	ER	M	CN	CS	ER
Original	0.55	4	256	0.50	0.68	4	256	0.15	0.67	8	256	0.50	0.80	8	256	0.15	0.68	6	256	0.33
RN	0.77	4.6	14.0	0.15	0.80	4.3	15.9	0.07	0.84	2.4	21.7	0.08	0.88	4.1	26.4	0.02	0.82	3.8	19.5	0.08
REN	0.62	6	14.0	0.15	0.72	**4.0**	50.0	0.03	0.73	**8.0**	50.2	0.17	0.85	**8.0**	50.4	0.02	0.73	6.2	48.4	0.10
RW	0.59	**4.2**	**48.2**	0.20	0.57	4.4	48.0	**0.20**	0.70	**8.0**	51.5	**0.19**	0.74	6.0	68.2	**0.03**	**0.65**	5.6	**54.0**	0.16
RJ	0.60	4.9	41.5	0.22	**0.69**	**4.0**	**50.5**	0.03	0.72	**8.0**	51.0	0.16	**0.83**	**8.0**	51.0	0.02	**0.71**	6.2	48.5	0.11
FF	**0.56**	4.9	41.8	**0.27**	0.45	6.5	33.5	0.62	**0.69**	7.5	**53.9**	0.17	0.66	5.0	**80.8**	**0.03**	0.59	**6.0**	52.5	**0.27**

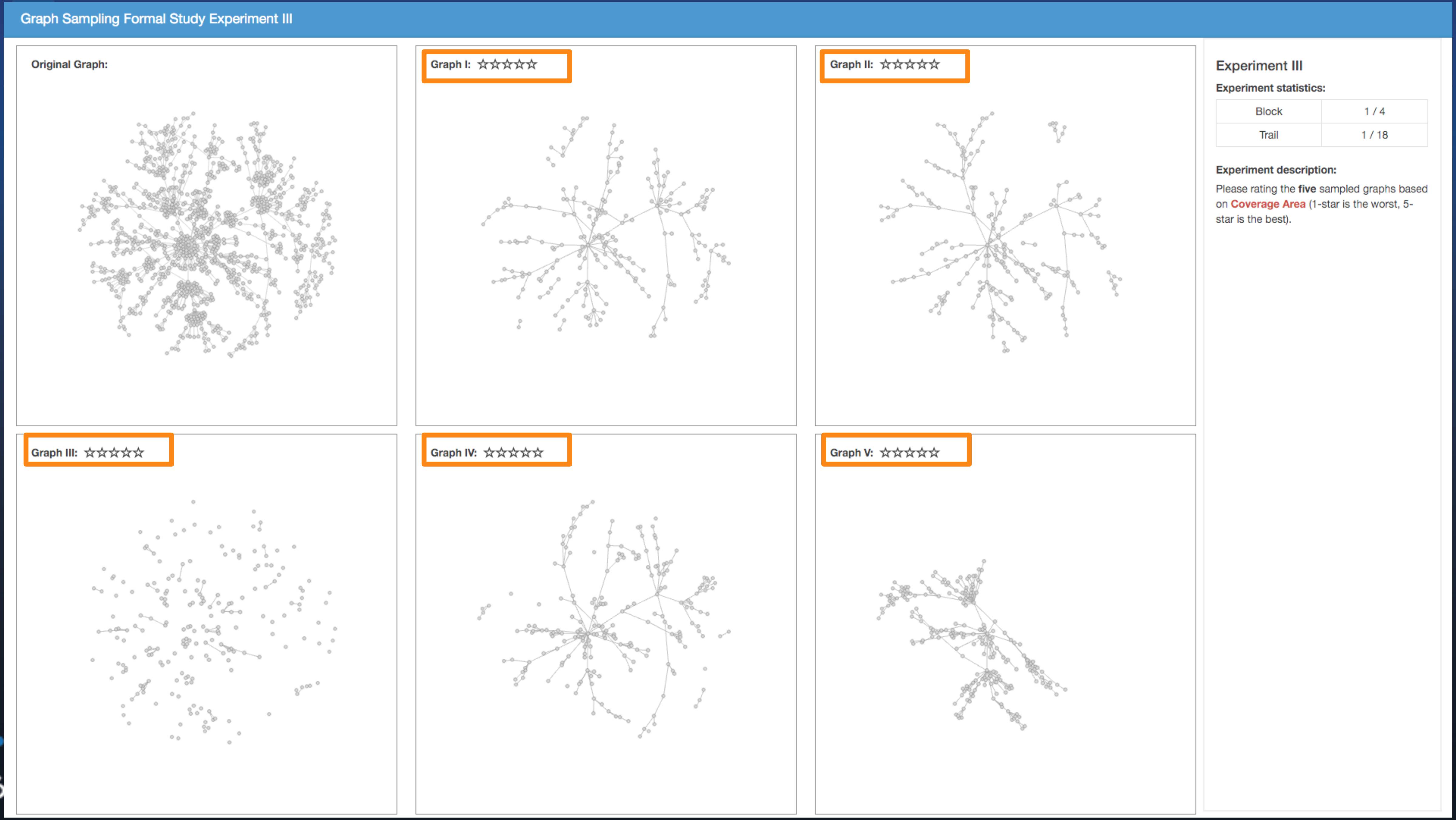


The number of clusters remained is important for perceiving the cluster quality in visualization!

Outline

- Selected Sampling Methods
- Pilot Study
- Formal Studies
  - Perception of High Degree Nodes (more high degree nodes are perceived in *RW*)
  - Perception of Cluster Quality (cluster number is important)
  - Perception of Coverage Area

Formal Study III: Coverage Area



Graph Sampling Formal Study Experiment III

Experiment III

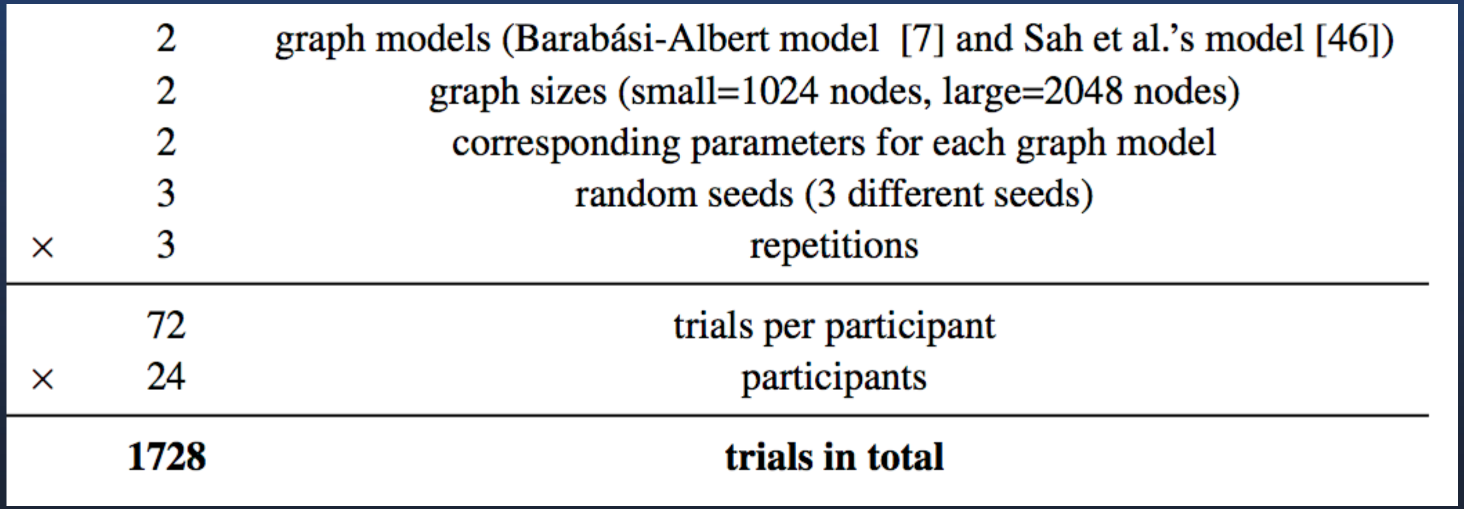
Experiment statistics:

Block	1 / 4
Trail	1 / 18

Experiment description:

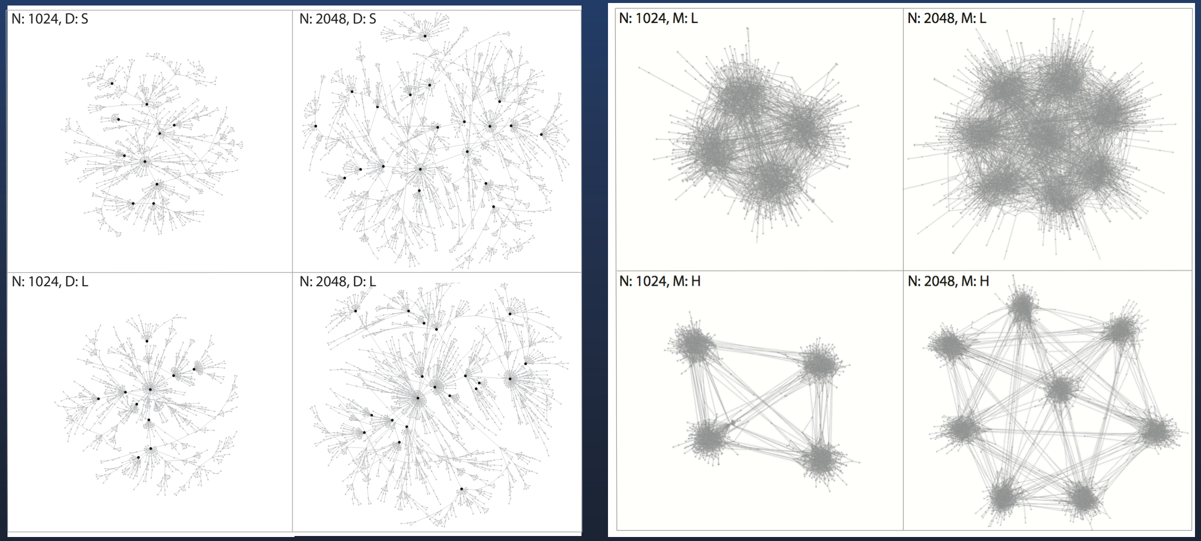
Please rating the **five** sampled graphs based on **Coverage Area** (1-star is the worst, 5-star is the best).

Formal Study III: Coverage Area



	2	graph models (Barabási-Albert model [7] and Sah et al.'s model [46])
	2	graph sizes (small=1024 nodes, large=2048 nodes)
	2	corresponding parameters for each graph model
	3	random seeds (3 different seeds)
×	3	repetitions
	72	trials per participant
×	24	participants
	1728	**trials in total**

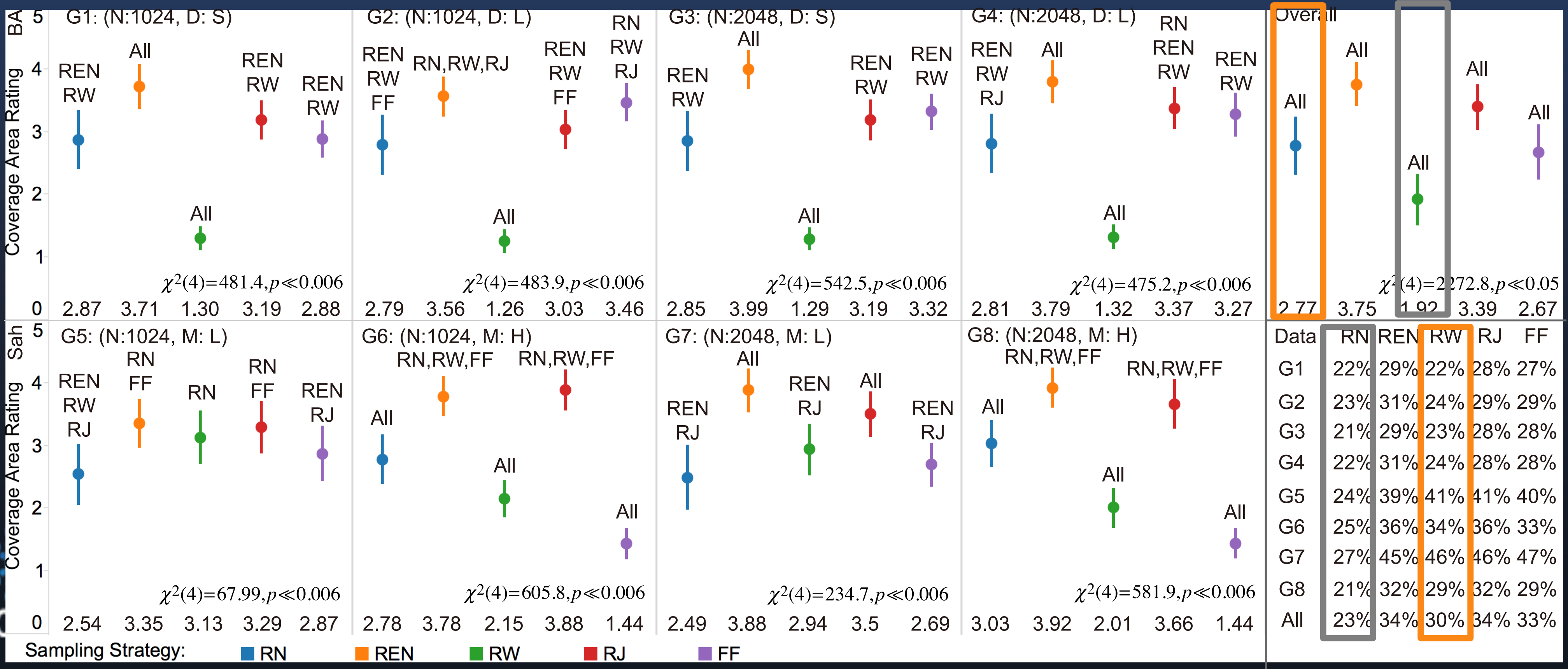
Experiment Setting



Data Generation

Formal Study III: Coverage Area Results

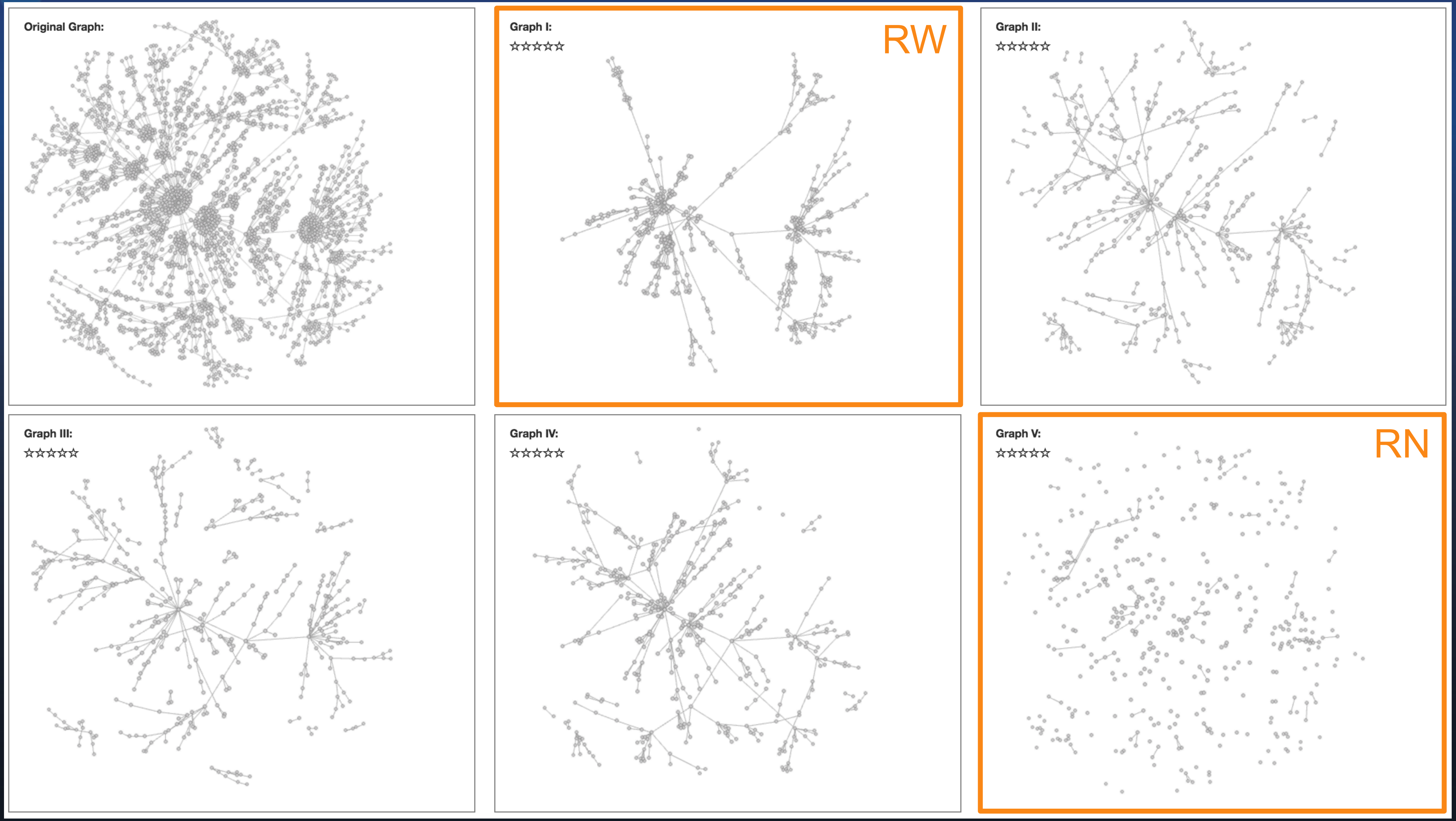
- Discussions:
  - *RE* and *RJ* have the largest perceived coverage area
  - *RW* has a smallest perceived coverage area in most cases
  - *RW* and *FF* 's performance vary depending on graph properties



Data	RN	REN	RW	RJ	FF
G1	22%	29%	22%	28%	27%
G2	23%	31%	24%	29%	29%
G3	21%	29%	23%	28%	28%
G4	22%	31%	24%	28%	28%
G5	24%	39%	41%	41%	40%
G6	25%	36%	34%	36%	33%
G7	27%	45%	46%	46%	47%
G8	21%	32%	29%	32%	29%
All	23%	34%	30%	34%	33%

Contradiction with metric-based results!

Formal Study III: Coverage Area Results

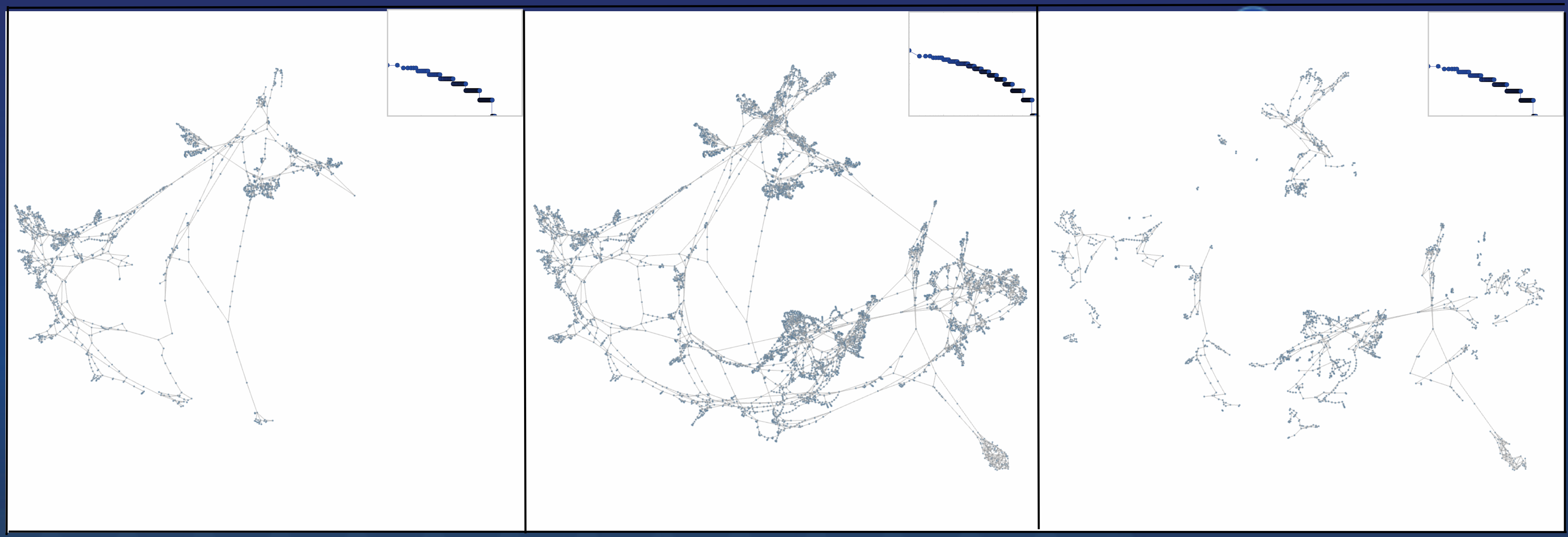


Conclusion

- We provided the first study of how graph sampling strategies can influence the perception of node-link visualizations
  - Important visual factors: high degree nodes, cluster quality, and coverage area
  - Recommendations for sampling network visualizations:
    - Recommend *Random Edge* and *Random Jump* for global structure and cluster quality
    - Recommend *Random Walk* for perceived high degree nodes
    - Use *Random Node* unless for specific requirements
    - *Random Walk* and *Forest Fire* are modularity sensitive

Graph sampling performance in visualization may **VARY** from previous metric-based results!

Q&A



Evaluation of Graph Sampling: A Visualization Approach

Yanhong Wu, Nan Cao, Daniel Archambault, Qiaomu Shen, Huamin Qu, and Weiwei Cui

yanhong.wu@ust.hk
http://yhwu.me

香港科技大學
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

上海纽约大学
NYU SHANGHAI

Swansea University
Prifysgol Abertawe

Microsoft Research
微软亚洲研究院