Centrality Preservation in Anonymized Social Networks

Traian Marius Truta[1], **Alina Campan**[1], Ashley Gasmi[2],
Nicholas Cooper[1], Andrew Elstun[1]

[1] Northern Kentucky University, USA
[2] ENSICAEN, France

Content of the Talk

- **Introduction**
- Social Network Privacy Model
- SaNGreeA Algorithm
- Graph Measures
- Experiments & Results
- Conclusions

Privacy in Social Networks

- Social networks tend to gather individuals' confidential information and/or confidential relationships between individuals.
  - Usual social tools such as Facebook
  - Specialized networks: PatientsLikeMe, Rareshare, Daily Strength, social networks in the healthcare field that create communities of patients for various diseases
- Consequently, privacy in social networks has become a serious concern for the large public and an active research field.

Privacy in Social Networks

- Identity and confidential information individual nodes of a social network should be protected in all situations.
- Anonymization of social network data and / or structure → a solution for privacy preservation in social networks
  - To anonymize a social network = to modify social network data and structure such that to make several individuals in the network alike, data and neighborhood-wise.
  - Several anonymity definitions and anonymization methods exist
    - Aim to preserve as much as possible the data and structural content of the initial social network.
    - Results obtained by exploring the anonymized social network – more accurate if social network is less "disturbed" in the anonymization process.

Privacy in Social Networks

- Contribution: our work studies how an existing anonymization approach preserves the structural content of the initial social network:
    - How various graph metrics (centrality measures, radius, diameter etc.) preserve through anonymization.
    - Study was performed for a number of synthetic social network datasets.

Content of the Talk

- Introduction
- **Social Network Privacy Model**
- SaNGreeA Algorithm
- Graph Measures
- Experiments & Results
- Conclusions

Social Network as a Graph

- We use the social network anonymization model from “Data and Structural K-Anonymity in Social Networks,” A. Campan and T. M. Truta, LNCS, vol. 5456, pp. 33-54, 2009.
- An undirected graph $\mathcal{G} = (\mathcal{N}, \mathcal{E})$,
  - $\mathcal{N}$ is the set of nodes
  - $\mathcal{E} \subseteq \mathcal{N} \times \mathcal{N}$ is the set of edges.
- Each node represents an individual entity.
- Each edge represents a relationship between two entities.

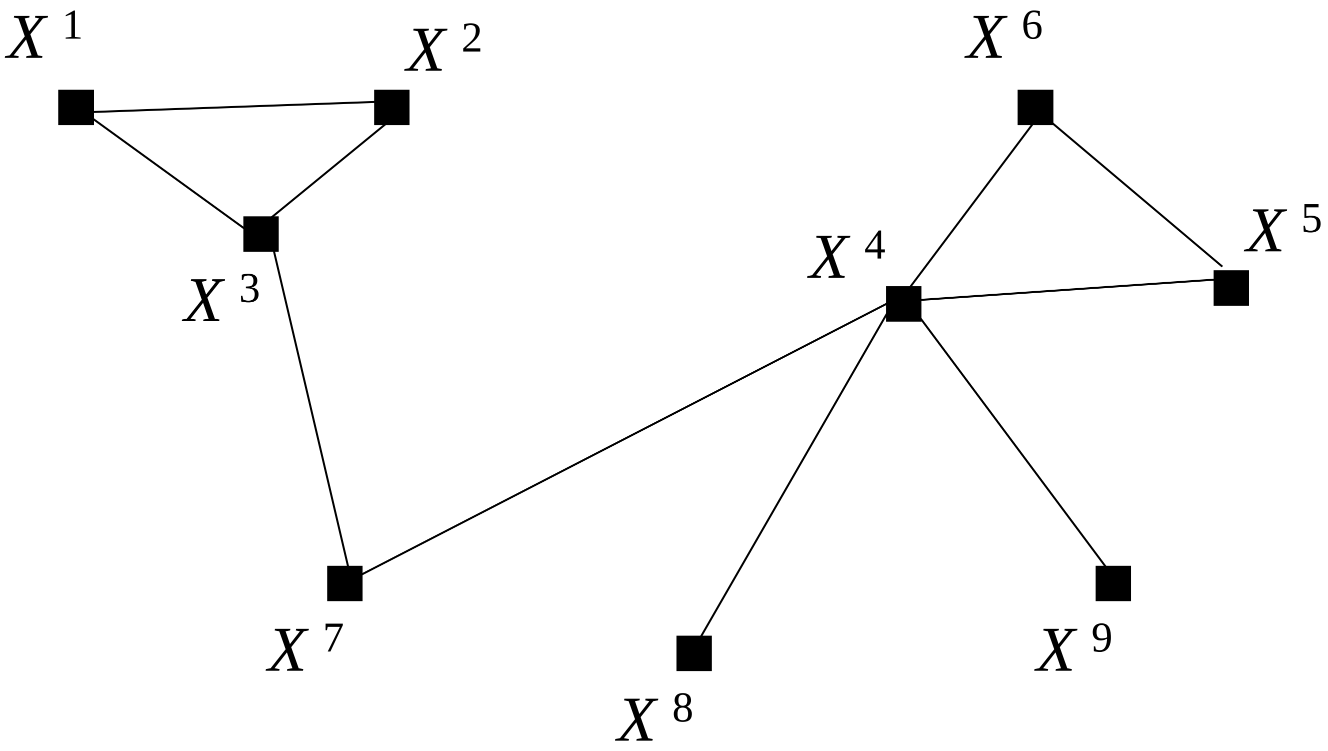
Node Attributes

- Nodes have several types of attributes, which have to be considered during anonymization, *BUT*
- We focus now only on social network structure and disregard node attribute values during the anonymization process.

Graph Edges

- Model binary relationships only.
- One type of relationship (unlabeled).
- We consider this structure to be of “quasi-identifier” type.
  - = the graph structure may be known to an intruder and used by matching it with known external structural information, therefore serving in attacks that might lead to identity and/or attribute disclosure
- We refer to this relationship as ***the quasi-identifier relationship***.

Running Example - 1



Privacy Model for Social Networks

- ***K-anonymity*** like model
  - Using a grouping strategy, one can partition the nodes from set $\mathcal{N}$ ($n=|\mathcal{N}|$) into v totally disjoint clusters: $cl_1, cl_2, \ldots, cl_v$.
  - Our goal is that any two nodes from any cluster to be indistinguishable based on both their attributes and relationships.
- ***Node generalization*** process – not discussed here
- ***Edge generalization*** process
  - *edge intra-cluster generalization*
  - *edge inter-cluster generalization*

Edge Intra-Cluster Generalization

- Given a cluster ***cl***, let $\mathcal{G}_{cl} = (cl, \mathcal{E}_{cl})$ be the subgraph of $\mathcal{G} = (\mathcal{N}, \mathcal{E})$ induced by *cl*.

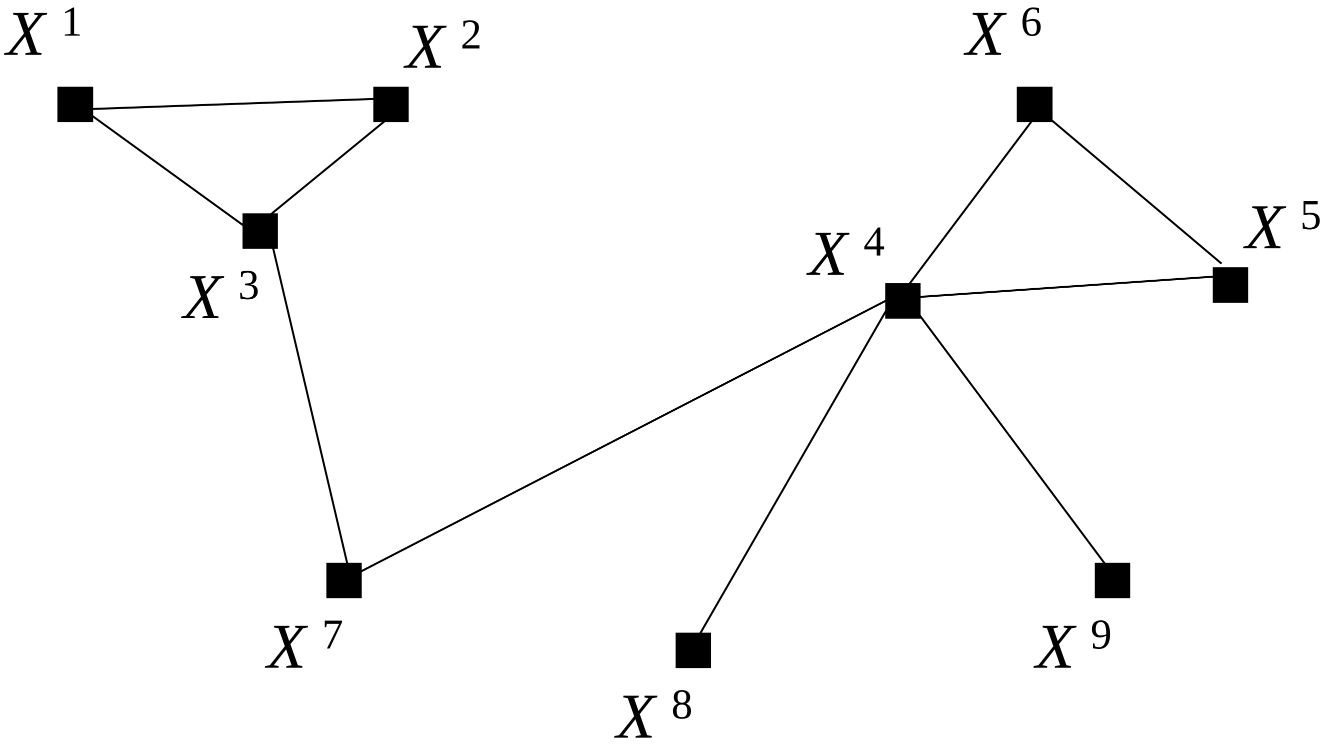
- In the masked data, the cluster ***cl*** will be generalized to (collapsed into) a node, and the structural information we attach to it is the pair of values $(|cl|, |\mathcal{E}_{cl}|)$, where $|x|$ represents the cardinality of the set ***x***.

Edge Inter-cluster Generalization

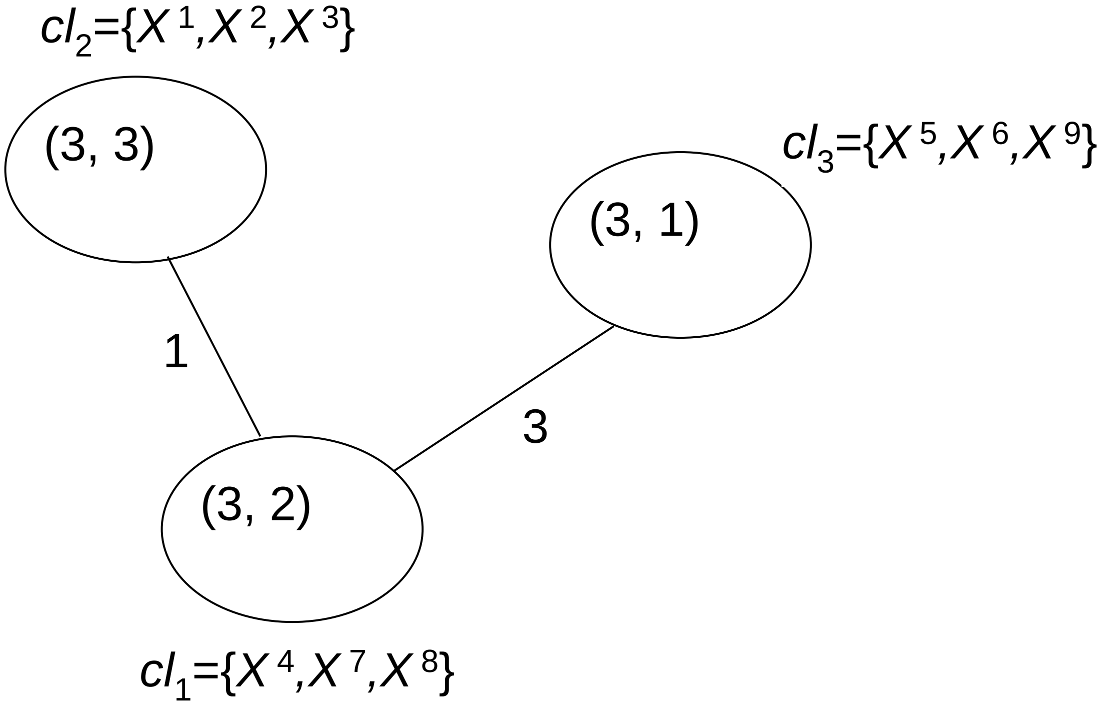
- Given two clusters ***cl_1*** and ***cl_2***, let $\mathcal{E}_{cl1,cl2}$ be the set of edges having one end in each of the two clusters ($e \in \mathcal{E}_{cl1,cl2}$ iff $e \in \mathcal{E}$ and $e \in cl_1 \times cl_2$).

- In the masked data, this set of inter-cluster edges will be generalized to (collapsed into) a single edge and the structural information released for it is the value $|\mathcal{E}_{cl1,cl2}|$.

Running Example - 2



Running Example - 3



K-Anonymous Masked Social Network

- Given a social network $\mathcal{G} = (\mathcal{N}, \mathcal{E})$, and a partition $\mathcal{S} = \{cl_1, cl_2, \dots, cl_v\}$ of the node set $\mathcal{N}$, the corresponding ***anonymized social network*** $\mathcal{AG}$ is defined as $\mathcal{AG} = (\mathcal{AN}, \mathcal{AE})$, where:
  - $\mathcal{AN} = \{Cl_1, Cl_2, \dots, Cl_v\}$; Cl_i is a node for the cluster $cl_j \in \mathcal{S}$, described by the intra-cluster generalization pair $(|cl_j|, |\mathcal{E}_{clj}|)$;
  - $\mathcal{AE} \subseteq \mathcal{AN} \times \mathcal{AN}$; $(Cl_i, Cl_j) \in \mathcal{AE}$ iif $Cl_i, Cl_j \in \mathcal{AN}$ and $\exists X \in cl_j$, $Y \in cl_j$, such that $(X, Y) \in \mathcal{E}$.

    Each generalized edge $(Cl_i, Cl_j) \in \mathcal{AE}$ is labeled with the inter-cluster generalization value $|\mathcal{E}_{cli,clj}|$.

- The anonymized social network $\mathcal{AG} = (\mathcal{AN}, \mathcal{AE})$, is ***k-anonymous*** iff $|cl_j| \geq k$ for all $j=1,\dots,v$.

Content of the Talk

- Introduction
- Social Network Privacy Model
- **SaNGreeA Algorithm**
- Graph Measures
- Experiments & Results
- Conclusions

Anonymization Algorithm

- *SaNGreeA* (Social Network Greedy Anonymization) algorithm, performs a greedy clustering processing to generate a *k*-anonymous masked social network.
- *SaNGreeA* puts together in clusters, nodes that are as similar as possible in terms of their neighborhood structure.

Anonymization Algorithm

- Proximity assessment of two nodes' neighborhood structures: we measure the degree to which the nodes have the same connectivity properties = are connected / disconnected among them & with others in the same way.
- Assume nodes in $\mathcal{N}$ have a particular order, $\mathcal{N} = \{X^1, X^2, \ldots, X^r\}$.
- The neighborhood of each node X^i is represented as an n-dimensional boolean vector, $B_i = (b_1^i, b_2^i, \ldots, b_r^i)$
  - b_j^i = 1 if there is an edge $(X^i, X^j) \in \mathcal{E}$, $\forall j = 1, r$; $j \neq i$
    = 0 if there is no edge $(X^i, X^j) \in \mathcal{E}$, $\forall j = 1, r$; $j \neq i$.
    = *undefined*, if $i=j$

Distance Functions

- ***Distance between two nodes*** = symmetric binary distance:

$$dist(X^i, X^j) = \frac{|\{\ell \mid \ell = 1..n \wedge \ell \neq i, j; b_\ell^i \neq b_\ell^j\}|}{n-2}$$

- ***Distance between a node and a cluster*** :

$$dist(X, cl) = \frac{\sum_{X^j \in cl} dist(X, X^j)}{|cl|}$$

SaNGreeA Algorithm

Algorithm *SaNGreeA* is

Input $\mathcal{G} = (\mathcal{N}, \mathcal{E})$ – a social network

k – as in k-anonymity

Output $\mathcal{S} = \{cl_1, cl_2, \ldots, cl_v\}$;

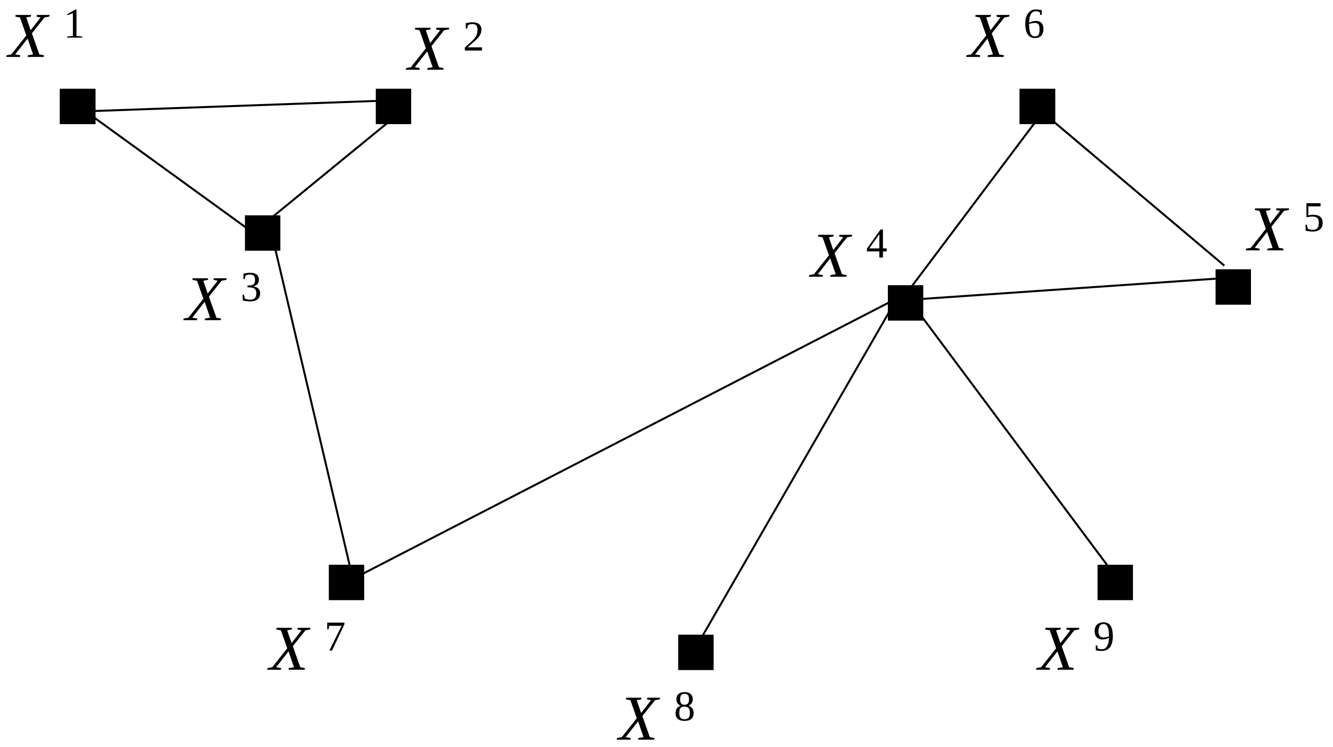
$\bigcup_{j=1}^{v} cl_j = \mathcal{N}$; $cl_i \cap cl_j = \varnothing$, $i,j=1..v$, $i \neq j$; $|cl_j| \geq k$, $j=1..v$ -

a set of clusters that ensures k-anonymity;

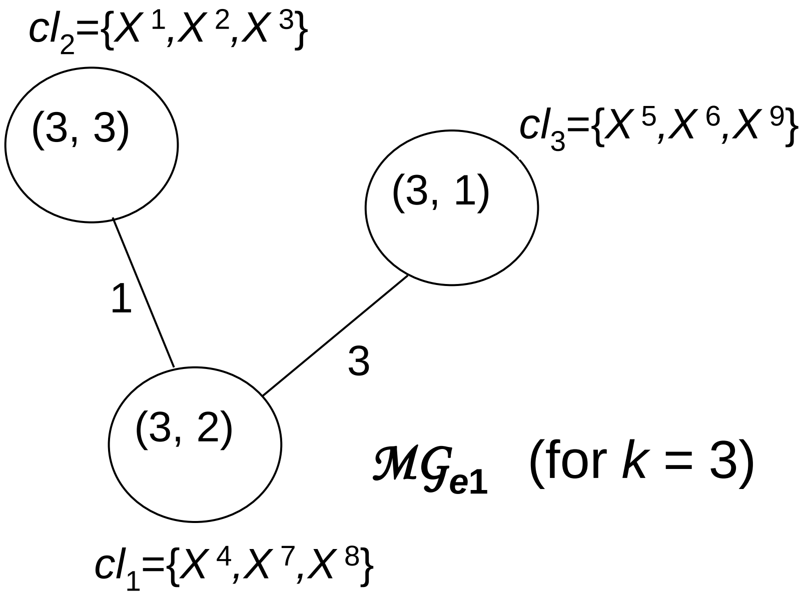
SaNGreeA Algorithm

```
$\mathcal{S} = \emptyset$; $i = 1$;
Repeat
  $X^{seed}$ = a node with maximum degree from $\mathcal{N}$;   $cl_i = \{X^{seed}\}$;
  $\mathcal{N} = \mathcal{N} - \{X^{seed}\}$; // $\mathcal{N}$ keeps track of nodes not yet distributed to clusters
  Repeat
    $X^* = \underset{X \in N}{\arg\min}(\, dist(X, cl_i))$
    // $X^*$ – a yet unselected node that produces a minimal IL growth when added to $cl_i$
    $cl_i = cl_i \cup \{X^*\}$;  $\mathcal{N} = \mathcal{N} - \{X^*\}$;
  Until ($cl_i$ has $k$ elements) or ($\mathcal{N} == \emptyset$);
  If ($|cl_i| \le k$) then DisperseCluster(S, $cl_i$);
      // This happens only for the last cluster: each of its nodes is added to the cluster
      // that is closest to that node w.r.t. our previously defined distance measure.
  Else
    $\mathcal{S} = \mathcal{S} \cup \{cl_i\}$; $i$++;
  End If;
Until $\mathcal{N} = \emptyset$;
End SaNGreeA.
```

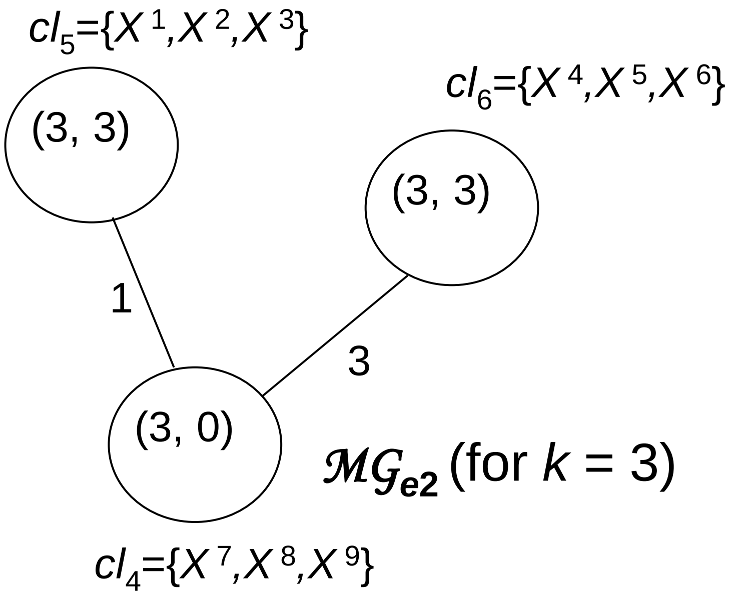
Running Example - 4



Running Example - 5


cl_2={X^1,X^2,X^3}
(3, 3)
cl_3={X^5,X^6,X^9}
(3, 1)
1
3
(3, 2)
cl_1={X^4,X^7,X^8}

$\mathcal{MG}_{e1}$ (for k = 3)


cl_5={X^1,X^2,X^3}
(3, 3)
cl_6={X^4,X^5,X^6}
(3, 3)
1
3
(3, 0)
cl_4={X^7,X^8,X^9}

$\mathcal{MG}_{e2}$ (for k = 3)

intraSIL	interSIL	SIL
$intraSIL(cl_1) = 4/3$ $intraSIL(cl_2) = 0$ $intraSIL(cl_3) = 4/3$	$interSIL(cl_1, cl_2) = 16/9$ $interSIL(cl_1, cl_3) = 4$ $interSIL(cl_2, cl_3) = 0$	$SIL(\mathcal{G},\mathcal{S}_1) = 8.444$
$intraSIL(cl_4) = 0$ $intraSIL(cl_5) = 0$ $intraSIL(cl_6) = 0$	$interSIL(cl_4, cl_5) = 16/9$ $interSIL(cl_4, cl_6) = 4$ $interSIL(cl_5, cl_6) = 0$	$SIL(\mathcal{G},\mathcal{S}_2) = 5.777$

Content of the Talk

- Introduction
- Social Network Privacy Model
- SaNGreeA Algorithm
- **Graph Measures**
- Experiments & Results
- Conclusions

Social Network Measures

- Graph connectivity and centrality metrics that quantify nodes' influence or power in the network.
- Connectivity:
  - radius
  - diameter
- Centrality:
  - degree centrality
  - betweenness centrality
  - closeness centrality

Social Network Measures

- Goal: explore the effect that social network anonymization has on various measures.
- Is there a relationship between these connectivity and centrality measures – for the initial social network and for a corresponding anonymized social network?
  - If the influence of a node on its network, as described by these measures, transferred from an original node to its supernode, then network analysis in various fields (viral marketing, communication networks) could be conducted on anonymized networks, while preserving the privacy of individual network nodes.

Social Network Measures – Connectivity

- Let $\mathcal{G} = (\mathcal{N}, \mathcal{E})$ be a social network, $|\mathcal{N}| = n$, $|\mathcal{E}| = m$.
- The **eccentricity of node *v*** is the maximum distance from *v* to any node:

$$\varepsilon(v) = \max\{d(v, w) \mid w \in \mathcal{N}\}.$$

- The **radius of $\mathcal{G}$** is the minimum eccentricity among the nodes of $\mathcal{G}$.

$$radius(\mathcal{G}) = \min\{\varepsilon(v) \mid v \in \mathcal{N}\}.$$

- The **diameter of G** is the maximum eccentricity among the nodes of G:

$$diameter(\mathcal{G}) = \max\{\varepsilon(v) \mid v \in \mathcal{N}\}.$$

SN Measures – Degree Centrality

- Nodes with more ties in the network have greater opportunities because they have choices ⇒ they are less dependent on any specific other node, therefore more powerful.

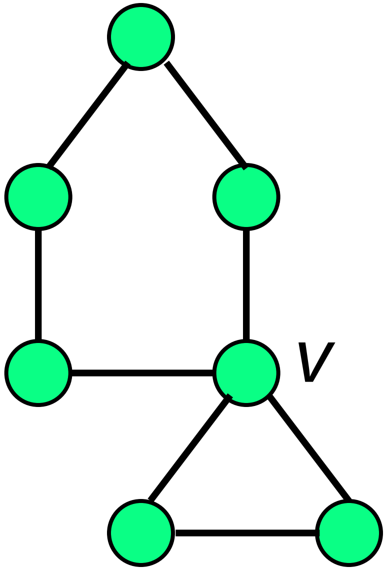
⇓

- The **degree centrality of node *v*** (communication potential) is the number of edges adjacent to the node (degree) normalized to the interval [0, 1]: $C_D(v) = \dfrac{deg(v)}{n-1}$

SN Measures – Degree Centrality

- Example:

  $C_D(v) = 4/6 = 0.67$



(From http://www.cs.umd.edu/~golbeck/CMSC498N/blog/3.2.pdf)

SN Measures – Betweenness Centrality

- Another aspect of a structurally advantaged position is in being between other nodes.
  - This gives a node the capacity to broker contacts among other nodes: to extract "service charges" and to isolate nodes or prevent contacts.

⬇

SN Measures – Betweenness Centrality

- The **betweenness centrality of node *v*** (potential for control of communication) is the sum of the number of shortest paths between any pair of vertices except *v*, going through *v*, divided by the number of shortest paths between any pair of vertices. This sum is normalized to [0, 1]:

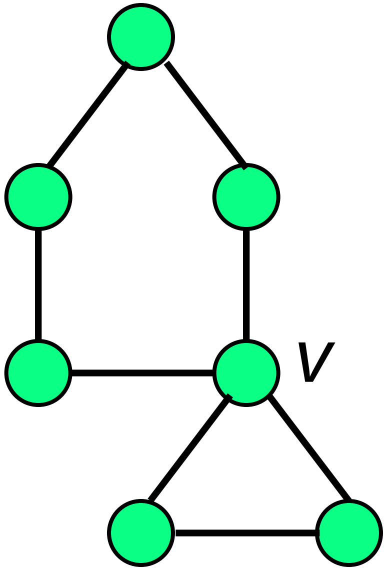
  $$C_B(v) = \frac{2 \cdot \sum_{s \neq v \neq t \in N} \frac{\sigma_{st}(v)}{\sigma_{st}}}{(n-1) \cdot (n-2)}$$

  where σ_{st} is the number of shortest paths from *s* to *t*, and $\sigma_{st}(v)$ is the number of shortest paths from *s* to *t* that pass through *v*.

SN Measures – Betweenness Centrality

- Example:

$C_B(v) = 2 \cdot 9/(49-21+2) = 9/15$



(From http://www.cs.umd.edu/~golbeck/CMSC498N/blog/3.2.pdf)

SN Measures – Closeness Centrality

- Nodes that are able to reach other nodes at shorter path lengths, or who are more reachable by other nodes at shorter path lengths have favored positions.

- The **closeness centrality of node *v*** (potential for independent communication) is the inverse of the average of shortest paths length between *v* and all other nodes from $\mathcal{G}$. This sum is normalized to [0, 1]:

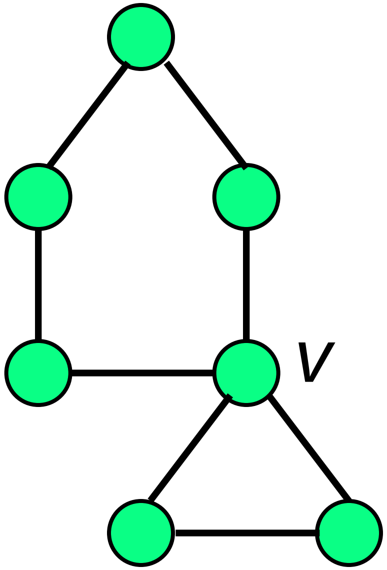
$$C_C(v) = \frac{n-1}{\sum_{i=1}^{n} d(v_i, v)}$$

where *d(v,w)* is the length of the shortest path from *v* to *w*

SN Measures – Closeness Centrality

- Example:

  $C_C(v) = 6/8 = 0.75$



(From http://www.cs.umd.edu/~golbeck/CMSC498N/blog/3.2.pdf)

SN Measures – Degree Centrality

- Note: power in social networks may be viewed both as:
  - a micro property (i.e. it describes relations between actors) *or*
  - a macro property (i.e. one that describes the entire population)

⇓

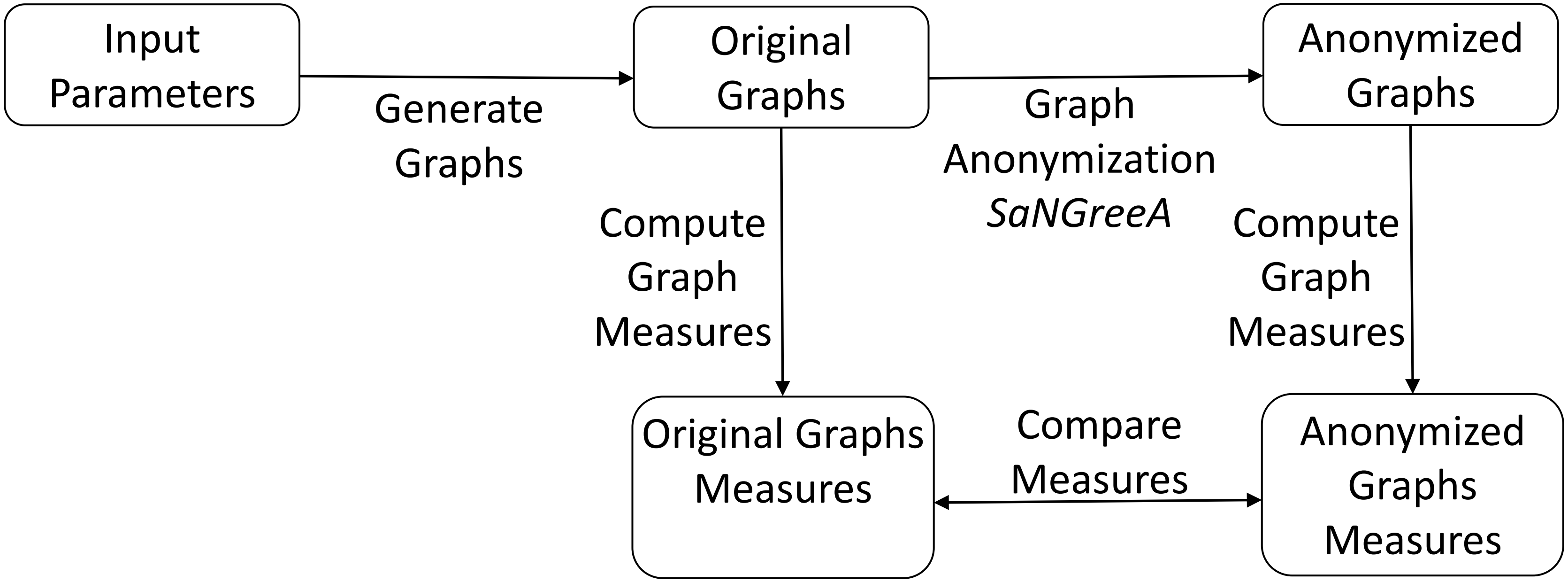
- Centrality measures are expressed both for individual nodes and for the entire network.
- The **degree centrality, betweenness centrality, and closeness centrality for a graph** $\mathcal{G}$ measure how much variation is there in the respective centrality scores among the nodes in $\mathcal{G}$.

Content of the Talk

- Introduction
- Social Network Privacy Model
- SaNGreeA Algorithm
- Graph Measures
- **Experiments & Results**
- Conclusions

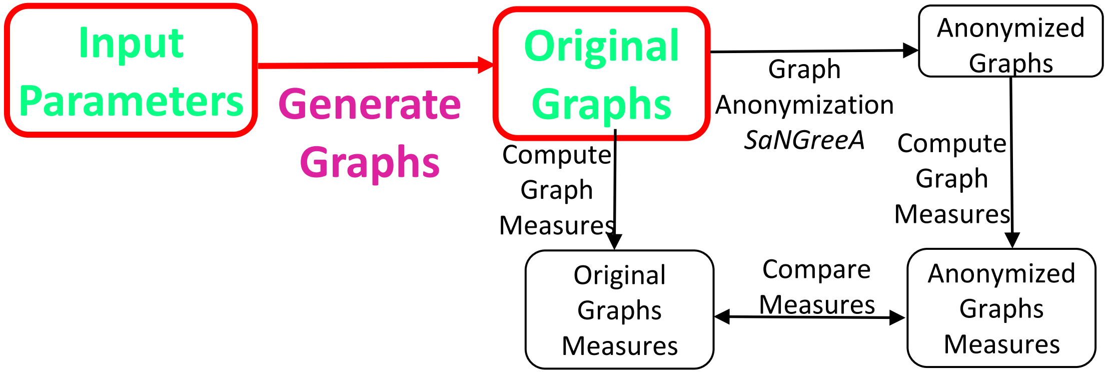
General Framework of the Experiments

- Design of experiments to empirically determine if the *SaNGreeA* graph anonymization algorithm preserves some of the graph properties, in particular centrality properties, of social networks.



Test Data

- Two graph generator models with various parameter values to create a large number of synthetic graphs on which we performed experiments:
  - R_MAT generator with parameters: number of nodes (*n*), average node degree (*avg_deg*), and 4 probabilities → we used 0.45, 0.15, 0.15, and 0.25 as values for the 4 probabilities, which seem to model better many real-world graphs that follow power-law degree distributions;
  - Random graph generator using the Erdos-Renyi model with 2 input parameters: number of nodes (*n*) and average node degree (*avg_deg*).



Input Parameters

Generate Graphs

Original Graphs

Graph Anonymization *SaNGreeA*

Anonymized Graphs

Compute Graph Measures

Compute Graph Measures

Original Graphs Measures

Compare Measures

Anonymized Graphs Measures

Test Data

- Parameter values:
  - *n* : 10, 25, 50, 75, 100, 250, and 500.
  - *avg_deg*: 2, 3, 4, 5, 8, 10, 25, 50, 75, 100, and 250.
  - *avg_deg* was strictly less than *n*-1 (no complete graphs).
- Most centrality measures are defined only for connected graphs.
  - ⇒ For every given combination of parameters we generated up to 10,000 graphs and we stopped the generator at the first connected graph.
  - In some cases (such as for *n* = 500, and *avg_deg* = 2) we were not able to generate a connected graph.

Test Data

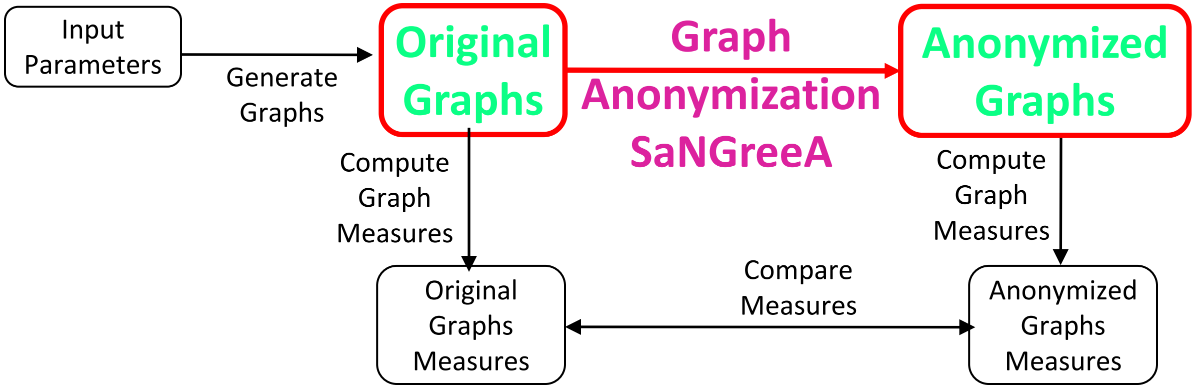
- The list of all generated graphs with the corresponding parameter values.

Graph Generator Model	(*n*, *avg_deg*)
R-MAT and RANDOM	(10, 2), (10, 3), (10, 4), (10, 5) (25, 2), (25, 3), (25, 4), (25, 5), (25, 8), (25, 10) (50, 3), (50, 4), (50, 5), (50, 8), (50, 10), (50, 25) (75, 4), (75, 5), (75, 8), (75, 10), (75, 25) (100, 4), (100, 5), (100, 8), (100, 10), (100, 25), (100, 50) (250, 5), (250, 8), (250, 10), (250, 25), (250, 50), (250, 100) (500, 8), (500, 10), (500, 25), (500, 50), (500, 100), (500, 250)

- The total number of generated graphs was 78.

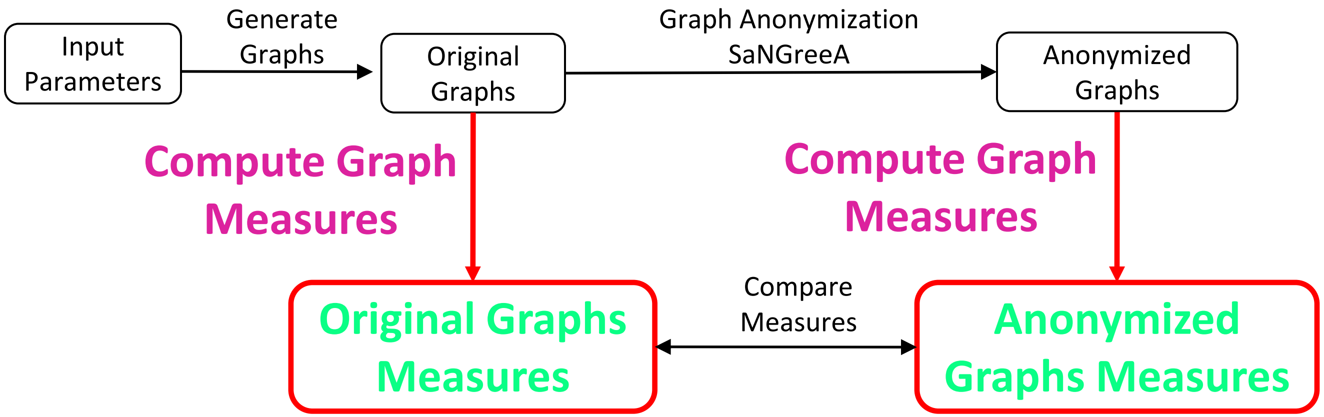
Anonymization

- For each generated graph we used various values for *k* (*k* as in *k*-anonymous social network).
  - For $n = 10$: $k = 2$ and 5;
  - For $n = 25$: $k = 2$, 5, and 10;
  - For all other values of n, $k = 2$, 5, 10, 15, and 20.
- In total 342 anonymized graphs were generated.



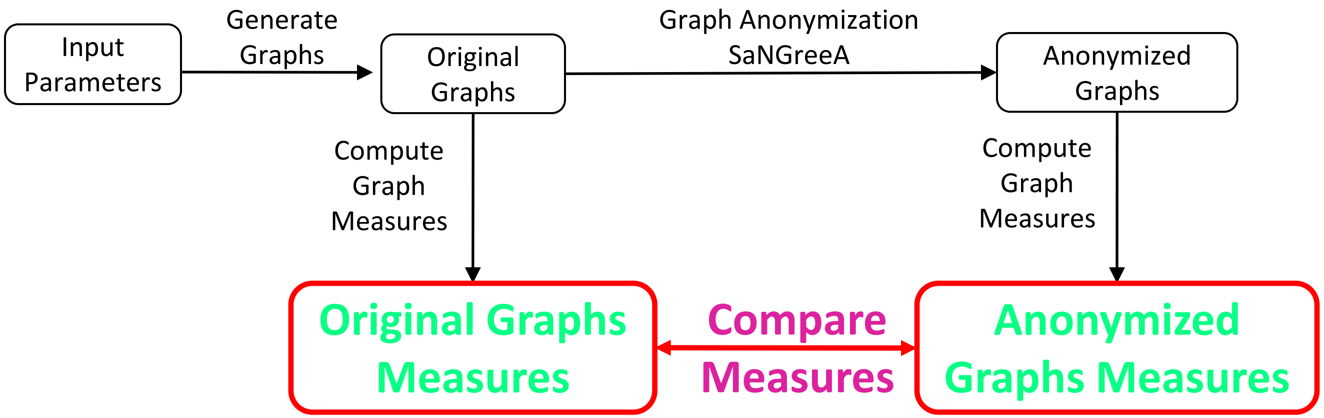
Graph Measures

- We implemented all graph measures described before.
- We computed these graph measures for all 420 graphs:
  - 78 original graphs *and*
  - 342 anonymized graphs.
- For an anonymized graph we did not use the weight of an edge between super-nodes, and we considered these graphs as unweighted graphs.



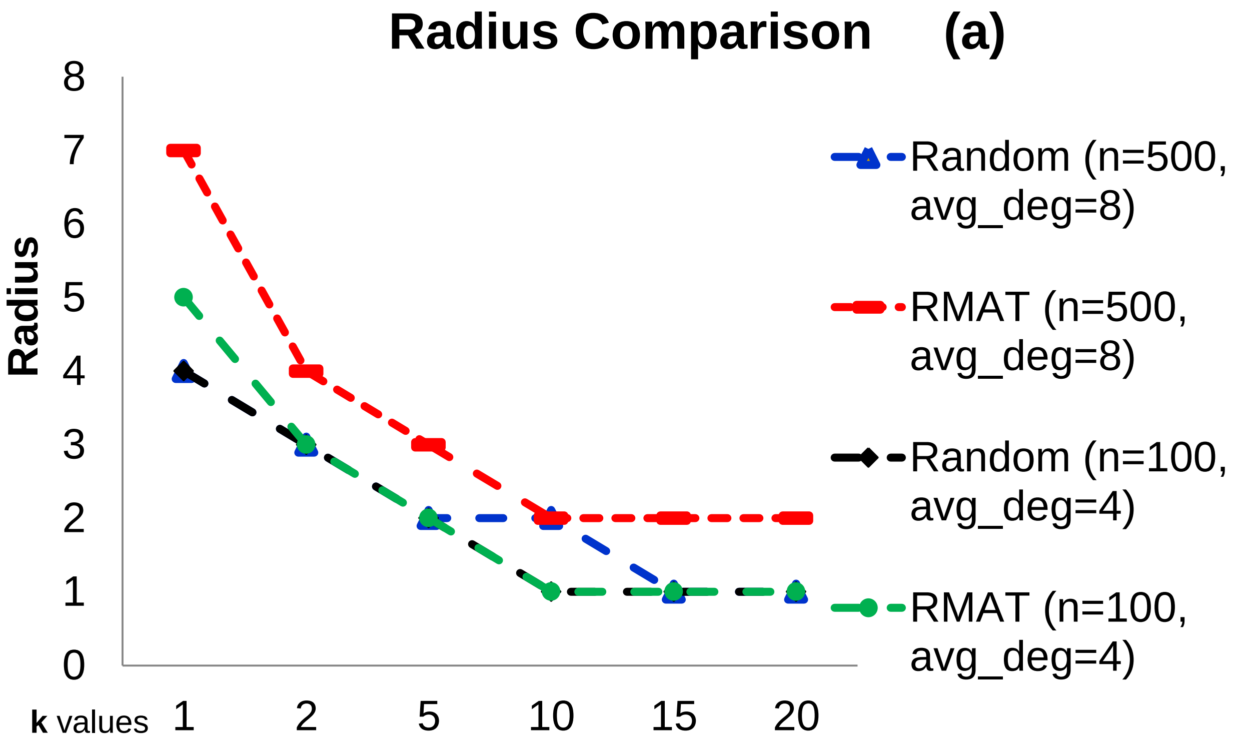
Experimental Results

- We are still in the process of analyzing results…



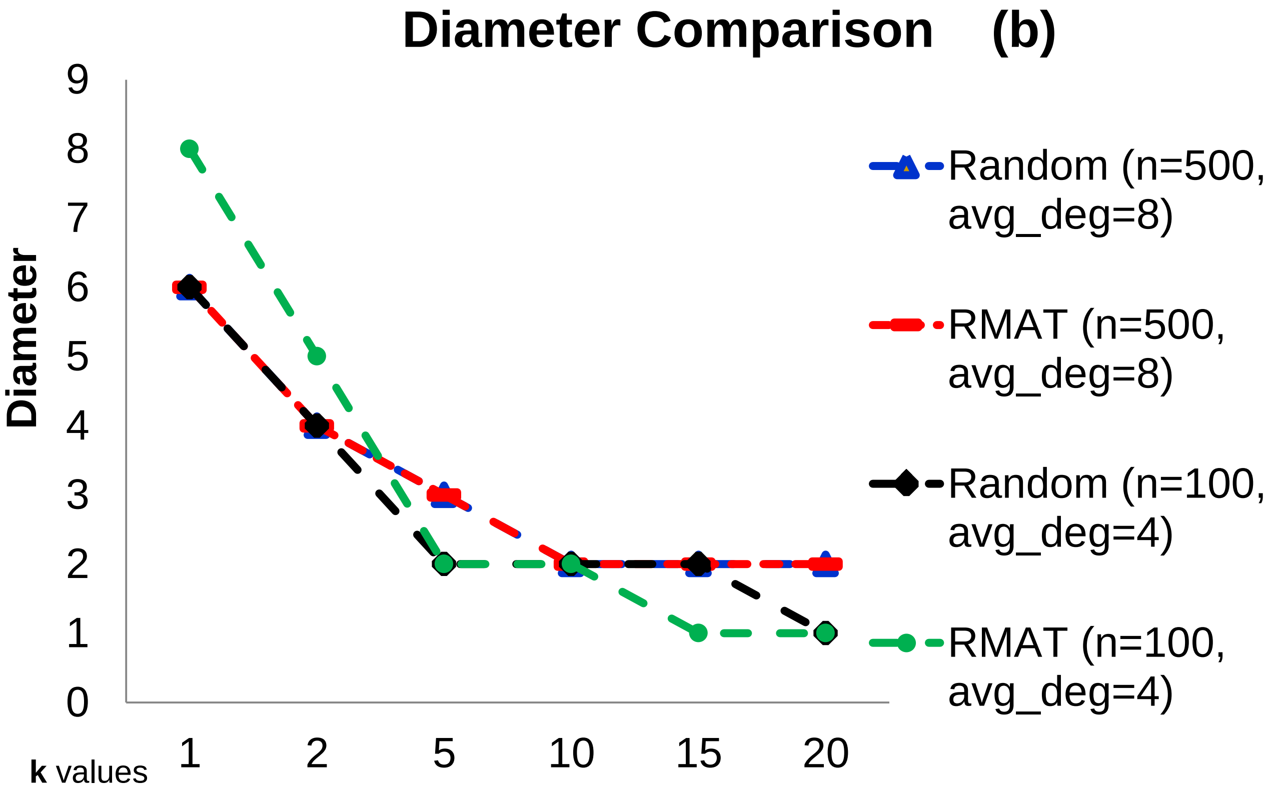
Radius and Diameter

- As expected, both these measures decrease as *k* increases.



Radius and Diameter

- As expected, both these measures decrease as *k* increases.



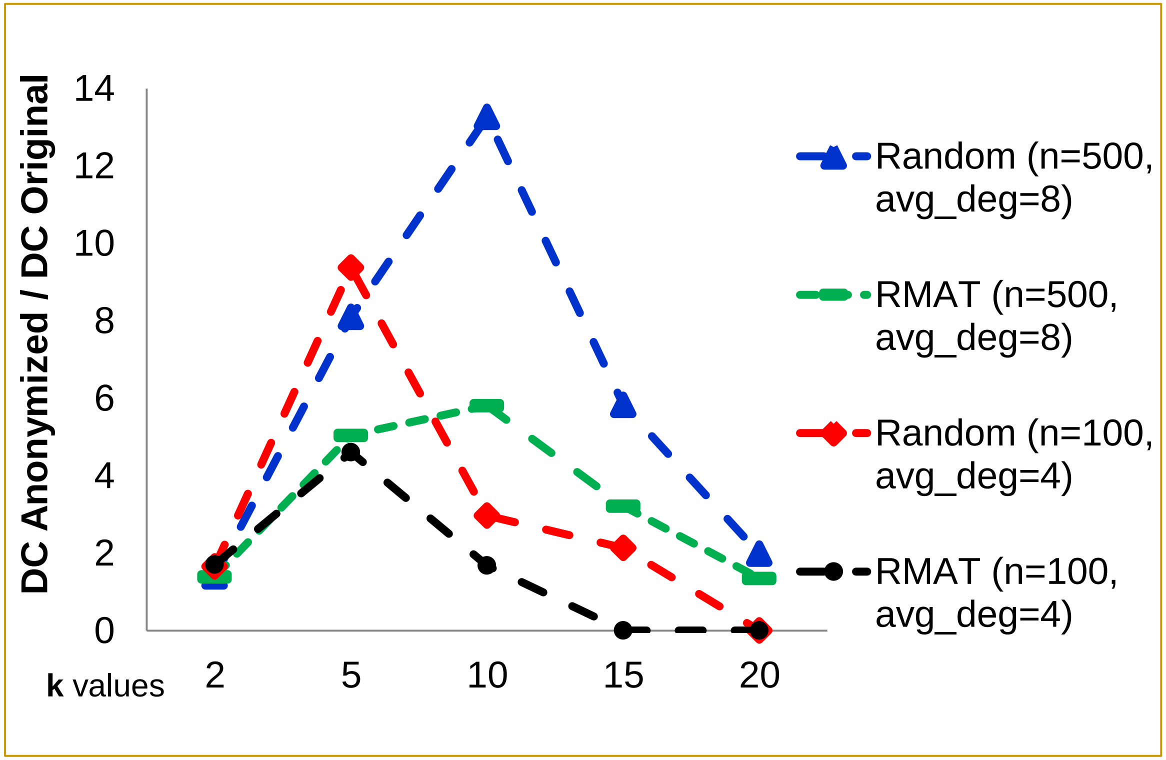
Centrality Measures

- For all measures we report the ratio between:
  - The measure value for the anonymized graph *and*
  - The measure value for the original graph.

  (⇒ The reference value for the original graph is 1 for all three measures).
- Results reported for 4 distinct original graphs:
  - 2 Random graphs and 2 RMAT graphs, with:
    - *n*=500 and *avg_deg*=8 (1 Random & 1 RMAT)
    - *n*=100 and *avg_deg*=4 (1 Random & 1 RMAT)
- For each original graph we created 5 *k*-anonymous graphs, for k ∈ {2, 5, 10, 15, 20}.

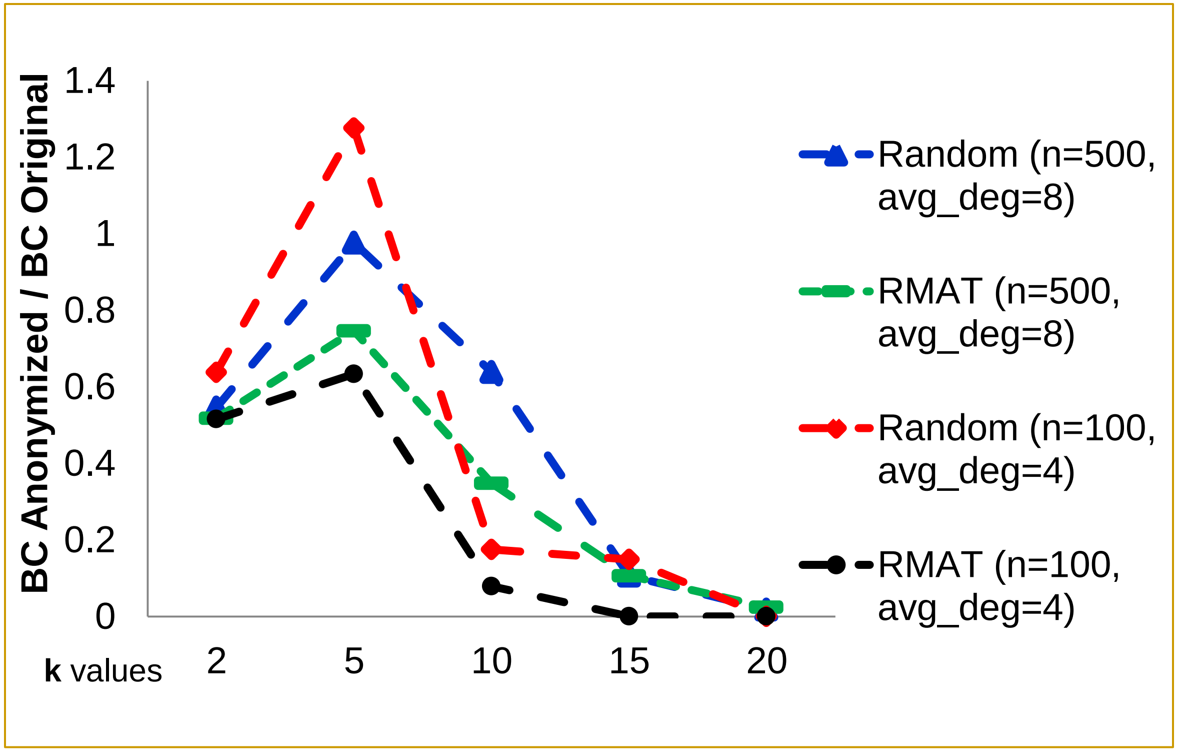
Degree Centrality

- DC increases as *k* increases to 5 / 10 (for smaller / larger graphs) and then decreases → due to how *SaNGreeA* creates clusters.
  - For small *k* values, supernodes created from nodes highly connected between them and loosely connected to other nodes
    ⇒ lower connectivity between supernodes
    ⇒ the anonymized graph is sparser than the original graph
  - For larger *k* values, supernodes made from nodes with different connectivity properties
    ⇒ the anonymized graph is closer to the complete graph
- Variation steeper for Random than RMAT - since original Random graphs have a uniform distribution of node degrees.



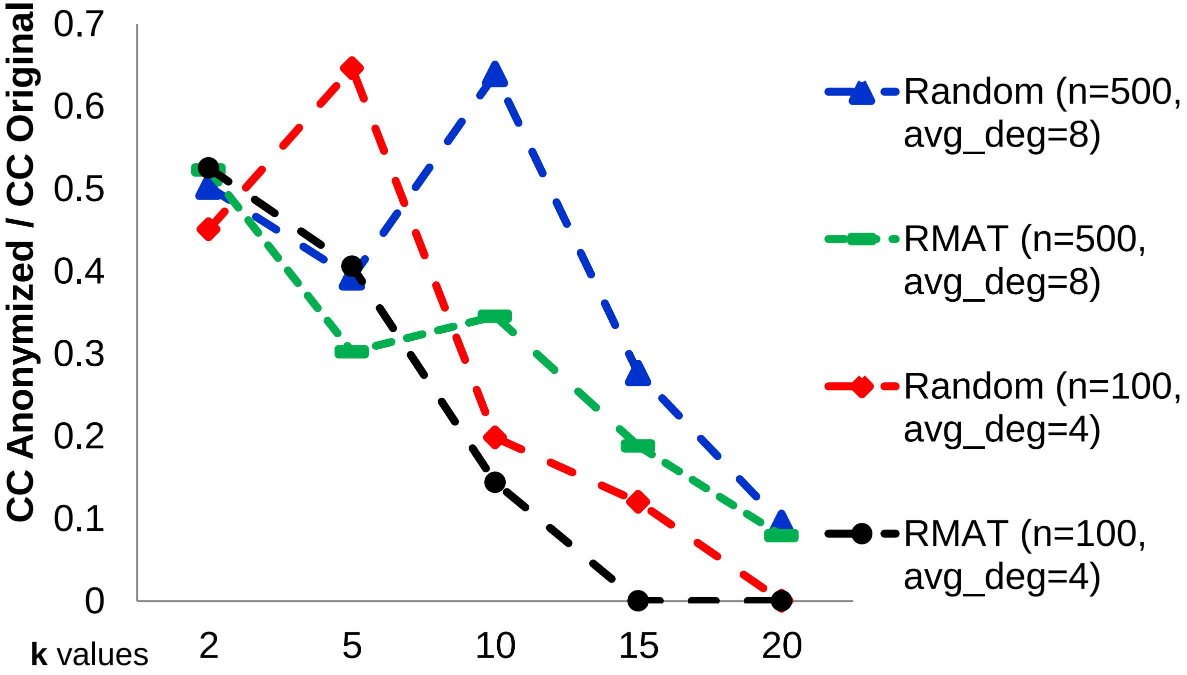
Betweenness Centrality

- BC usually decreases for the anonymized graphs. Reason:
  - The anonymized graph gets closer to the complete graph as *k* increases, ⇒ There are many short paths of length 1.
- A small increase between *k* = 2 and *k* = 5. Reason:
  - For small *k*s, the anonymized graph still has variety in supernodes' connectivity
    ⇒ Some supernodes gain more control over the shortest paths that exist in the anonymized graph;
    ⇒ These nodes have high betweenness centrality.



Closeness Centrality

- CC decreases for anonymized graphs when the value of *k* increases as shown in Figure.
  - This is again due to the anonymized graph getting closer to the complete graph.



Content of the Talk

- Introduction
- Social Network Privacy Model
- SaNGreeA Algorithm
- Graph Measures
- Experiments & Results
- **Conclusions**

Contributions and Future Work

- We studied a clustering-based anonymization approach with respect to how it preserves the structural content of the initial social network:
  - We looked at how various graph metrics (centrality measures, radius, diameter etc.) change between the initial and the anonymized social network.
  - Our experiments showed a weak correlation between the anonymization level (*k* value) of a graph and the centrality measures: same changes are observed for graphs of different sizes and with different network properties.
- We plan to study how other anonymization models behave with respect to centrality measures.

Questions

- For questions, comments, and suggestions, please contact me at:

campana1@nku.edu