CS511 Final Presentation

Team: Select * .....
Mentor: Jia Wang
Members: Bassi Bhopesh, Tubbs Dustyn James, Jingnan Yang, Tianhang Sun

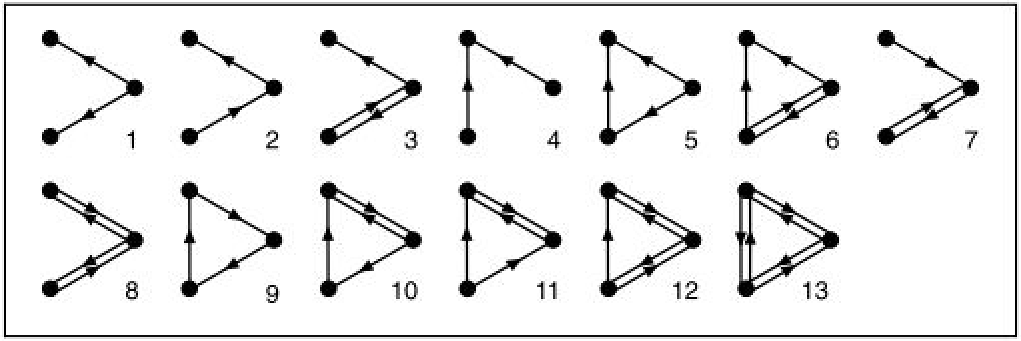
Project: General Social Pattern Discovery

- Problem: Conventionally, community detection tries to find subgraphs of a network that are densely connected
    - Sometimes the structure of a subgraph is as important as the connections themselves
    - Different types of networks can present interesting, and very different structures that define a "community"

- A better wording: How can we build a system that is flexible in the manner that lets one define what a community is, and what it is not?

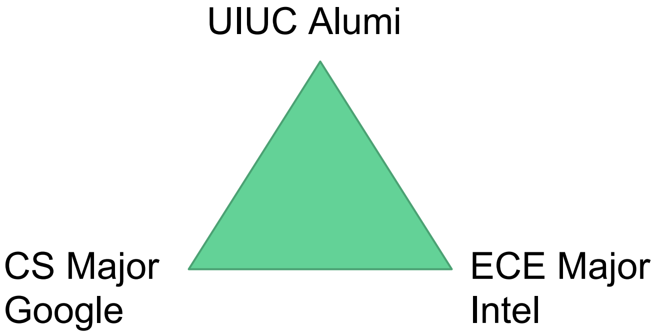
Our Approach - Graph Motifs

- “Recurrent, statistically significant subgraphs”
- Flexible structures that can be used to describe a subgraphs in a manner that is similar to a regular expression.
- Different network types typically demonstrate differing types of motifs found, helping describe the network itself



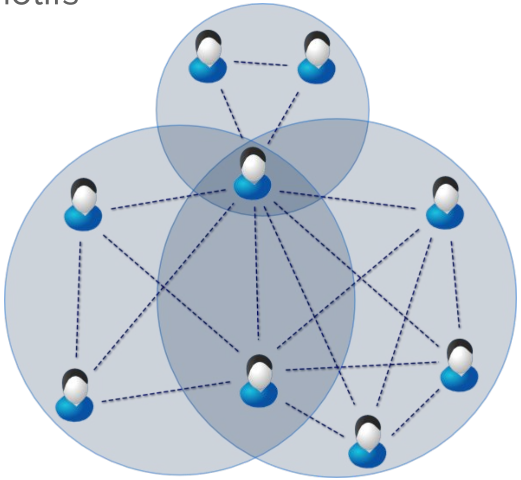
Our Approach - Generalizing K-Truss

- K-Truss is an algorithm that builds out networks for a very particular subgraph structure of “K” support
  - Wang, Jia, and James Cheng. "Truss decomposition in massive networks." *Proceedings of the VLDB Endowment* 5.9 (2012): 812-823.
- Our work: build a system that enables a user to define the community structure they’re looking for in the form of a typed motif
  - Types will denote attributes on the node



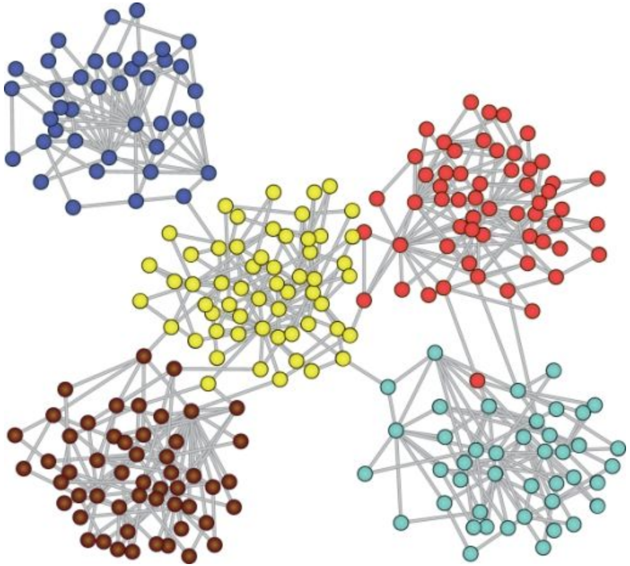
Our Approach - Generalizing K-Truss Cont.

- Generalize the process presented in K-Truss by building such “structurally, mutually supporting” networks
- Take the original structures and collapse them to represent a single node and relationships between these structures represented by an edge
  - The edge will denote the level of support between the two motifs



Our Approach - Generalizing K-Truss Cont.

- Cluster the resultant “motif graph” to find the strongest communities composed of the desired typed motif

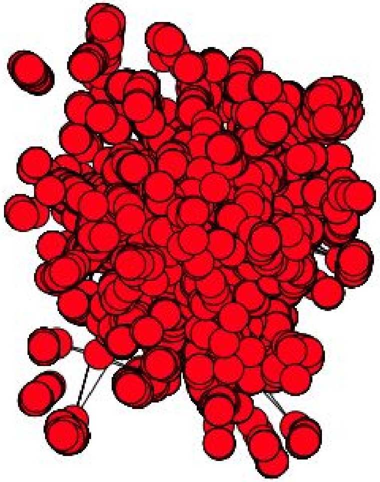


Overall Framework:

- Query all occurrence of typed (attributed) motifs in original network
  - We designed an algorithm base on the paper: Sun, Zhao, et al. "Efficient subgraph matching on billion node graphs." *Proceedings of the VLDB Endowment* 5.9 (2012): 788-799.
  - The algorithm in the paper only works on “labeled” graph, where each node can have only one label, we extend the algorithm to support multiple attributes.
- Construct a new motif network
  - Each node is a motif in original network
  - Edges represent the shared node in the motifs
- Clustering the newly constructed motif network
  - Existing library based on the paper: Newman, Mark EJ. "Fast algorithm for detecting community structure in networks." *Physical review* E 69.6 (2004): 066133.

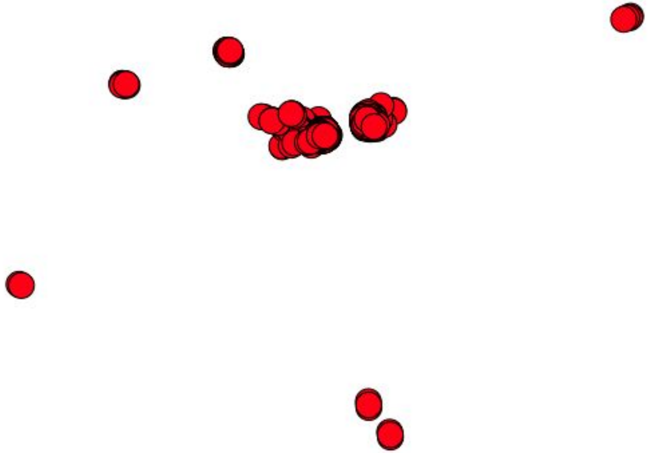
Results - Dataset / Clustering on Original Network

- Based on LinkedIn dataset
  - number of nodes: 29040, number of edges: 58080
  - nodes have attributes like employers, education, locations
  - no existing dataset with ground truth community labels based on attributed motifs
- Clustering on original network
  - clustering time: 2738 s, number of clusters: 77
  - the clusters found are shown on the right
  - is this meaningful ? probabaly yes, but?



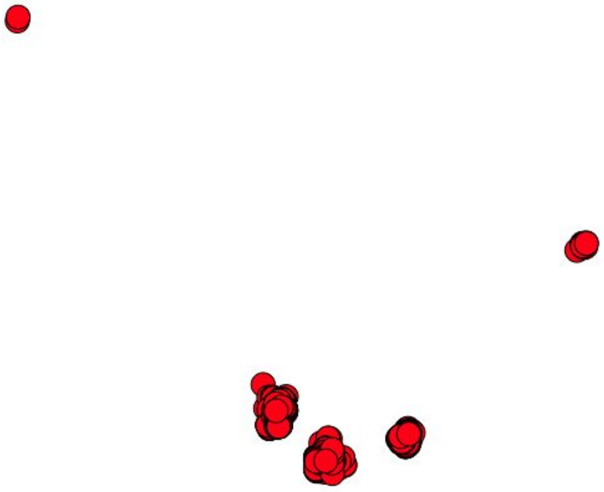
Results - Clustering Using Motif Network

- Clustering using motif network, using triangle motifs
  - (uiuc, google, facebook)
    - capture employee network at tech company in bay area that are uiuc alumni
    - motif query time: 0.208 s, network construction time: 0.505 s, occurrence of motifs: 1564, clustering time: 10.733 s, number of clusters: 11
    - considerably less communities
    - faster running time



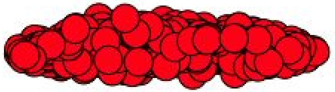
Results - Clustering Using Motif Network

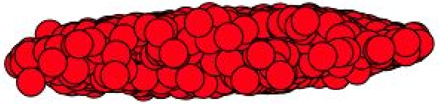
- Clustering using motif network, using triangle motifs
    - (uiuc, sjtu, google)
        - capture network of alumni from both uiuc/sjtu at google
        - motif query time: 0.260 s, network construction time: 0.473 s, occurrence of motifs: 1636, clustering time: 10.456 s, number of clusters: 5
        - sjtu is from China, thus fewer clusters



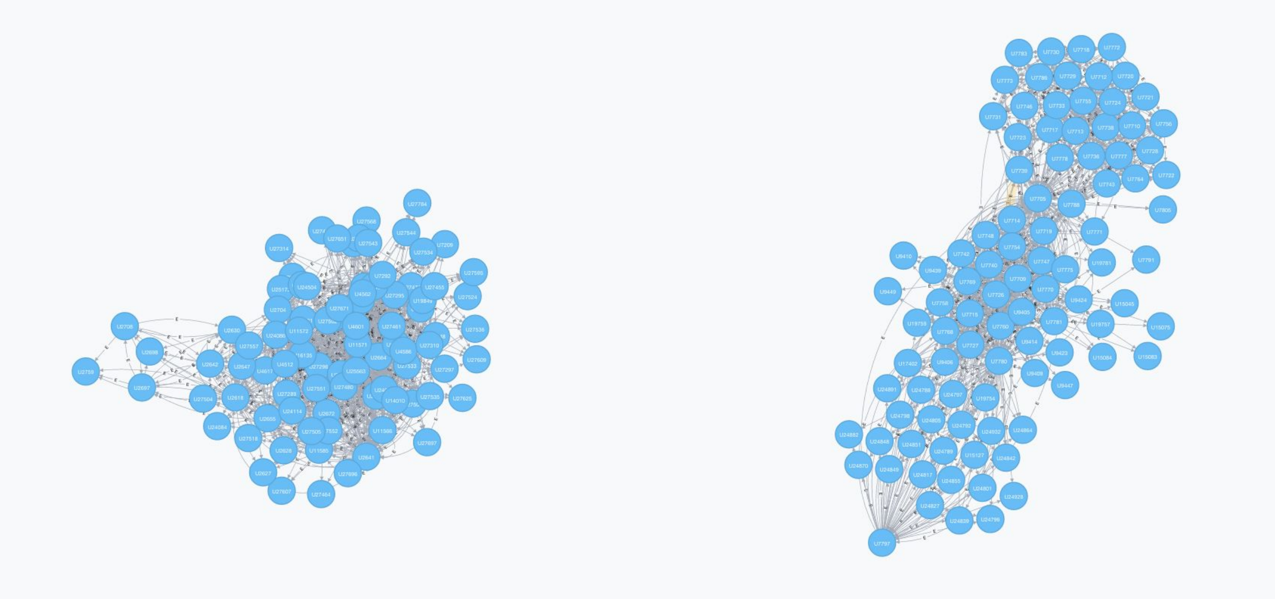
Results - Clustering Using Motif Network

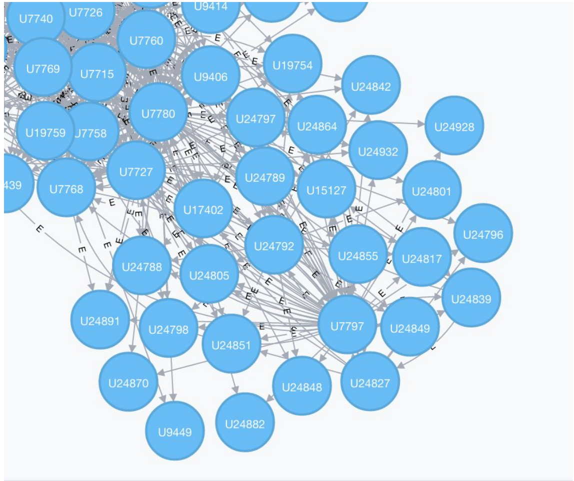
- Clustering using motif network, using star motifs
  - facebook -> [facebook, facebook, facebook, facebook]
    - capture facebook emplyee network
    - motif query time: 0.456 s, network construction time: 4.813 s, occurrence of motifs: 1753, clustering time: 38.832 s, number of clusters: 2
    - there’re fewer communities within a company
    - network is more densely connected



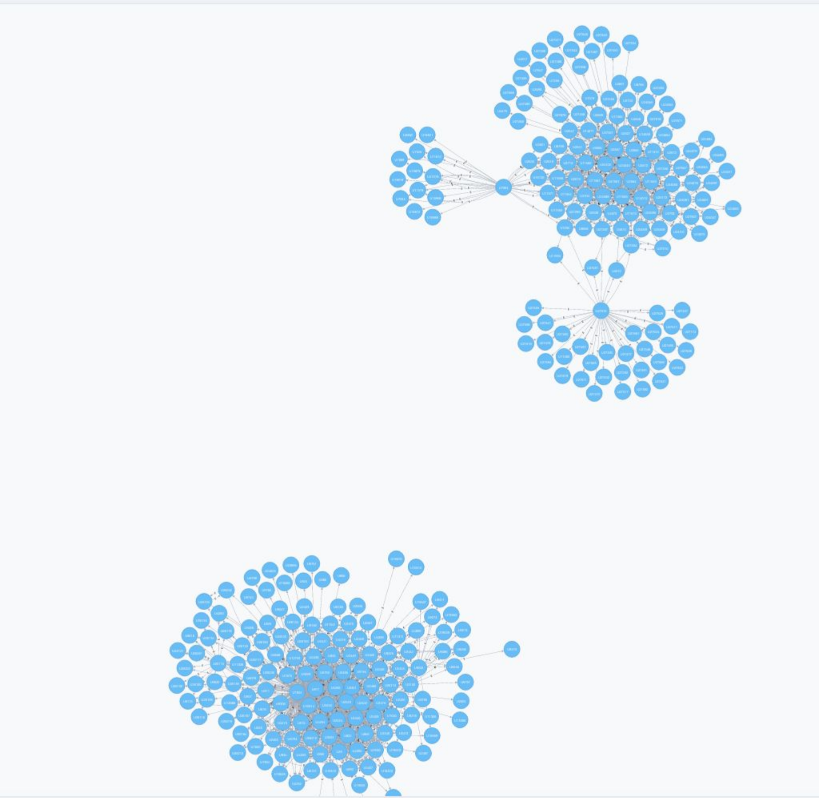


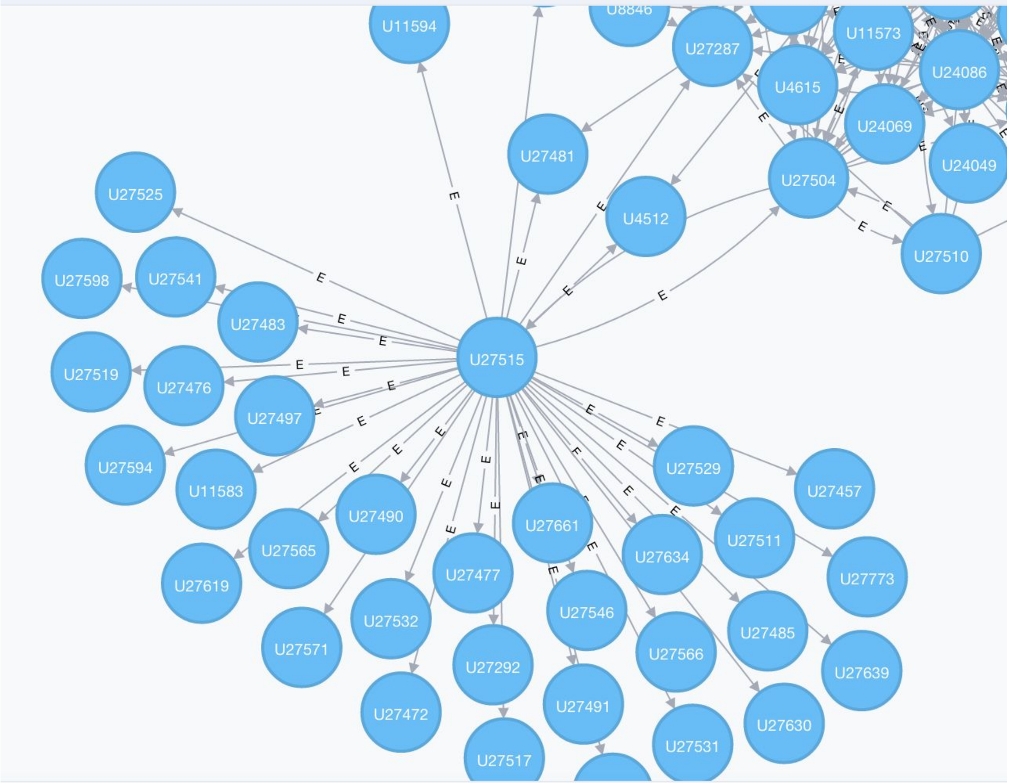
Results on Orignal Network - (uiuc, sjtu, google)





Results on Orignal Network - (fb -> [fb, fb, fb, fb])





Takeways

- By using motifs, we can find different, but interesting communities from using original network
- If the attribute is not occurring everywhere, using motif network can result in considerably faster running time
- Clustering algorithm complexity is the bottleneck in our framework

Future Work

- Make graph clustering algorithm aware of weights of connection, and faster
  - Our motif network construction takes into account the weight of the connection, but the graph clustering algorithm doesn’t work with weighted edges.
  - The algorithm runs in O(n^2), which couldn’t scale to very large graphs, eg. whole LinkedIn graph
- Optimize the motif discovery algorithm when motif is not triangle / star
  - Our motif discovery algorithm works for any motifs, but we only optimized for triangle / star motifs, when using other motifs, there’s a expensive joining phase which can be optimized, also, we could do motif exploration to further speed up the algorithm.
- Needs evaluation on a dataset which nodes have attributes and there’s ground truth communities based on the motifs
  - No such dataset exists (that we could find)

Thank You !