UNIVERSITY
OF
ALBERTA

ALBERTA INGENUITY CENTRE FOR MACHINE LEARNING

Top Leaders Community Detection Approach in Information Networks

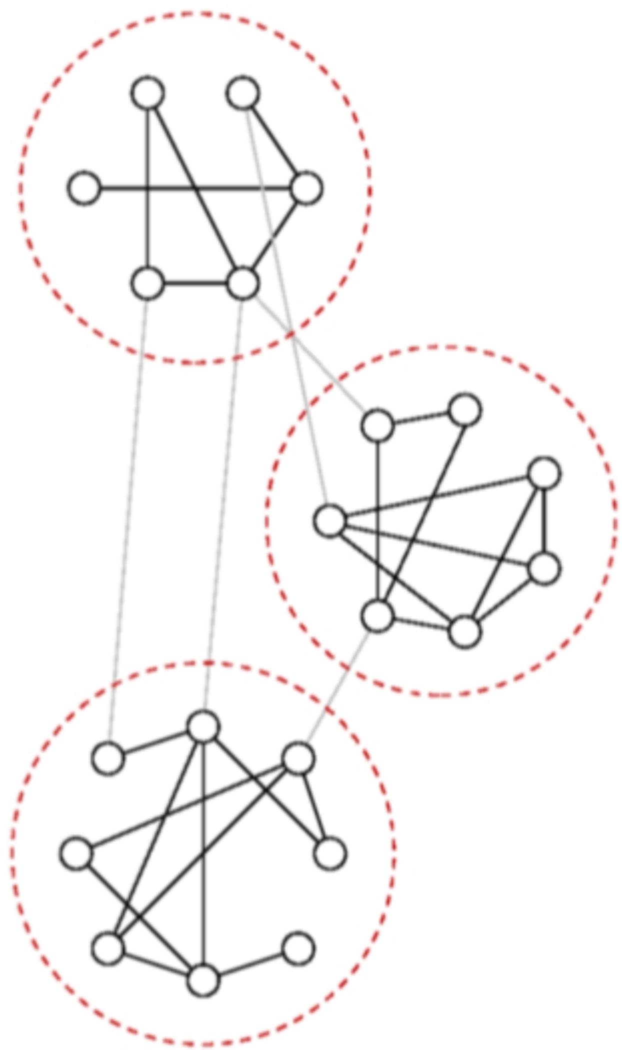
Reihaneh Rabbany Khorasgani, Jiyang Chen, Osmar R. Zaïane

Department of Computing Science
University of Alberta
Edmonton, Alberta, Canada T6G 2E8
{rabbanyk, jiyang, zaiane}@cs.ualberta.ca

4th SNA-KDD Workshop on Social Network Mining and Analysis – Washington DC, July 2010

What is Community Structure?

- *Community structure* denotes the existence of densely connected groups of nodes, with only sparser connections between groups.
- Many social networks share the property of a community structure, e.g., WWW, tele-communication networks, academic collaboration networks, friendship networks, etc.



Many similarities with data **Clustering**

Outline

- Introduction
- Some related work
  - Graph Partitioning, Modularity Q and others
- Top Leaders
  - Problem Definition
  - Associating nodes to Leaders
  - Initialization of the leaders
  - Experiments
- Conclusions

Introduction

- A new Approach for Finding Communities in the Networks
  - Densely connected nodes with sparse connections outside the group

- Community: set of followers congregating around a potential leader

- Algorithm: similar in spirit to k-means
  - starts by identifying promising leaders in the network
  - iteratively until convergence
    - assembles followers to their closest leaders to form communities
    - finds new leaders in each group around which to gather followers again

Related Work

- Graph partitioning and spectral clustering approaches
  - dividing the network into groups with (roughly) equal size, while minimizing the number of edges that run between vertices in different groups
- Q-modularity by Newman 2004
  - measure of the quality of a particular division of a network
- CFinder by Palla et al. 2005
  - community: union of complete subgraphs of size k
  - k between 3 and 5: very effective on real networks
- SCAN by Xu et al. 2007
  - nodes that are structurally reachable from each other are grouped together in the same community

Problem Definition

- Finding leaders and their followers in the network to form the communities
- Leader: The central/influential node
- Community: Set of followers surrounding a leader
- Assigning followers to closest leader based on the intersection of their neighborhoud

Top Leaders Approach

A leader is the most central member in a community

Algorithm 1 Top Leaders algorithm

Input: A social network G, and k the number of desired communities

```
initialize k leaders
repeat
    {finding communities}
    for all Node n ∈ G do
        if n ∉ leaders then
            associate n to a leader {Algorithm 2}
        end if
    end for
    {updating leaders}
    for all l ∈ leaders do
        l ← arg max_{n∈Community(l)} Centrality(n)
    end for
until there is no change in the leaders
```

Associating Nodes to Leaders

Algorithm 2 Associate n to its leader

Input: Social network G, node n, set of k leaders

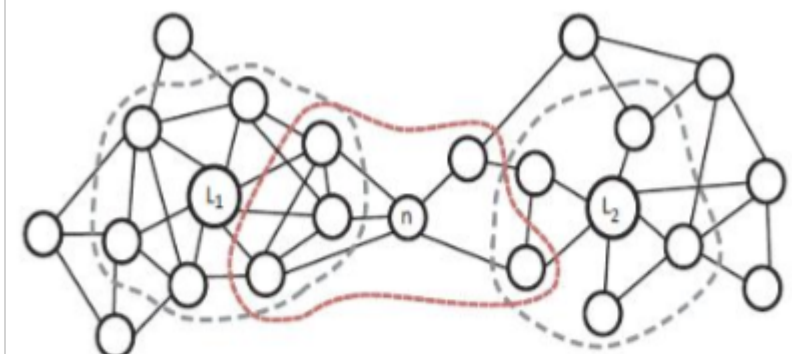
depth $\leftarrow$ 1
CanList $\leftarrow$ leaders
repeat

$$\text{CanList} \leftarrow \underset{\substack{c \in CandList \wedge \\ |\aleph(n_1,d) \cap \aleph(n_2,d)| > \gamma}}{\arg\max} |\aleph(n_1,d) \cap \aleph(n_2,d)|$$

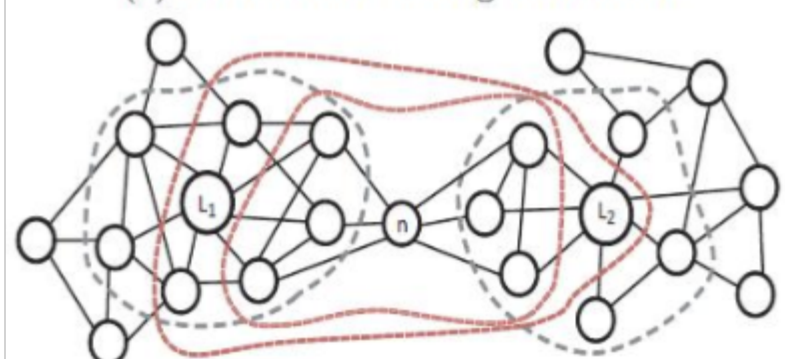
depth $\leftarrow$ depth+1
until $|\text{CanList}| \leq 1 \vee \text{depth} > \delta$

if $|\text{CanList}| = 0$ **then** {No candidate leader}
associate n as an outlier
else if $|\text{CanList}| > 1$ **then** {Many candidates}
associate n as a hub
else {Only one candidate leader in CanList}
associate n to CanList
end if

Community membership of the nodes is association of followers to nearby leaders



(a) Intersection of neighbourhoods



(b) Expanding Neighbourhoods

Top Leaders Approach

A leader is the most central member in a community

Algorithm 1 Top Leaders algorithm

Input: A social network G, and k the number of desired communities

initialize k leaders
repeat
  {finding communities}
  for all Node n $\in$ G **do**
    if n $\notin$ leaders **then**
      associate n to a leader {Algorithm 2}
    end if
  end for
  {updating leaders}
  for all l $\in$ leaders **do**
    l $\leftarrow \arg\max_{n \in Community(l)} Centrality(n)$
  end for
until there is no change in the leaders

Initialization Methods

Wrong leaders may get stuck in a bad local optimum

- Naïve Initialization
  - random selection of k nodes from the network
- Top Global Leaders
  - k most central nodes in the network
- Top Leaders & not Direct Neighbour
  - the k most central nodes that are not directly connected to each other
  - avoid choosing two correct leaders that are directly connected but truly in different communities
- Top leaders & Few Neighbours in Common
  - based on intersections, similar to followers association

Experiments and Datasets

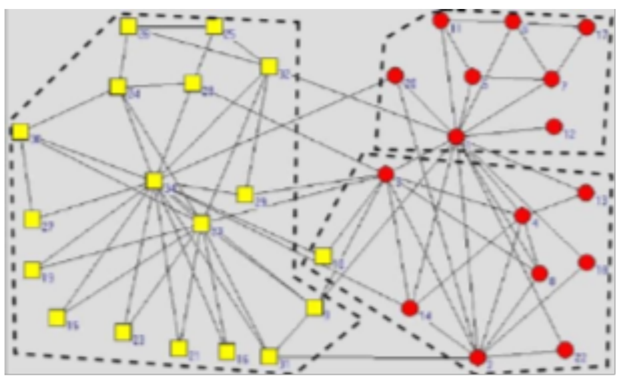
Competitors

- three of other well-known community detection methods
  - SCAN (KDD 2007), CFinder (Nature 2005) and FastModularity (2004)

Datasets

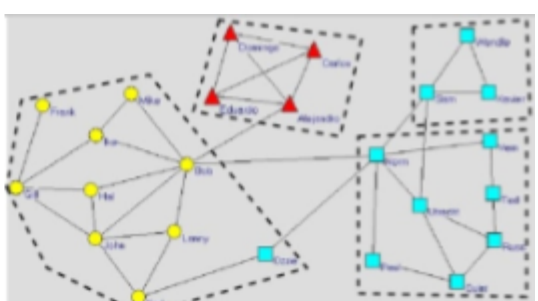
- Karate-Club dataset

34 nodes in 2 communities



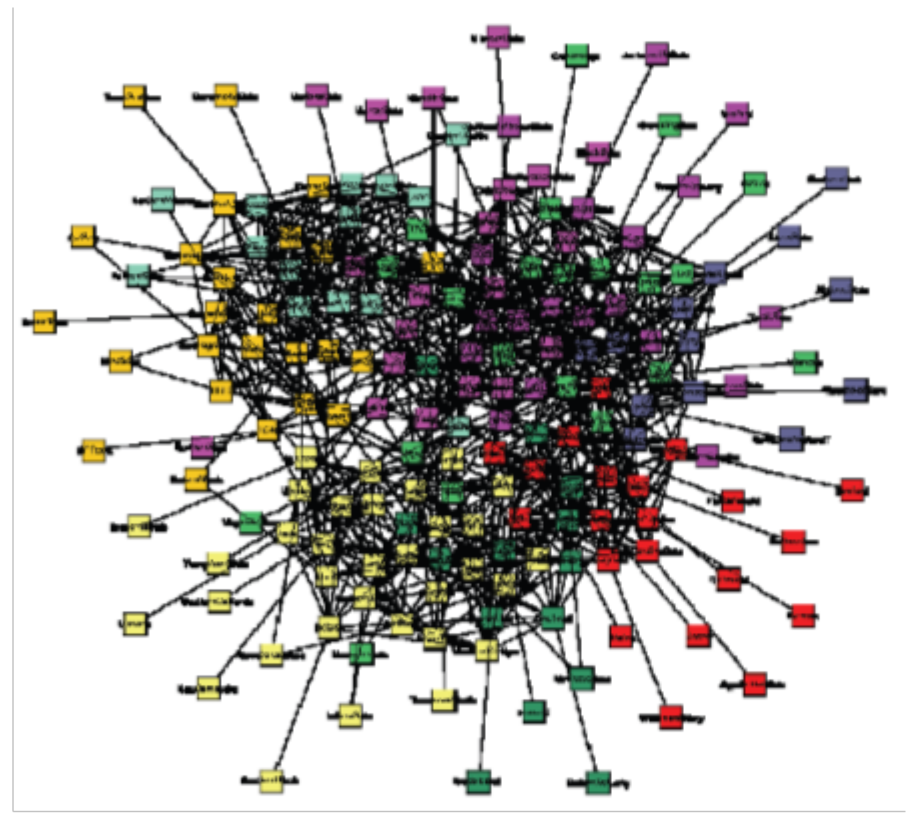
- Sawmill Strike dataset

24 nodes in 3 communities



- Football dataset

180 nodes in 11 communities



Evaluation Metrics

- Comparing with Ground Truth
  - Purity
    - the number of correctly assigned nodes divided by the total number of nodes. **0 (no agreement at all) to 1 (full agreement)**.
  - Adjusted Rand Index (ARI)
    - penalizes false negatives and false positives. **-1 (no agreement at all) and 1 (full agreement), 0 (no better than random)**

- Modularity
  - how well the edges fall within the detected communities compared to a randomized network. **0 (no different than a randomized network), > 0.3 (good partition)**

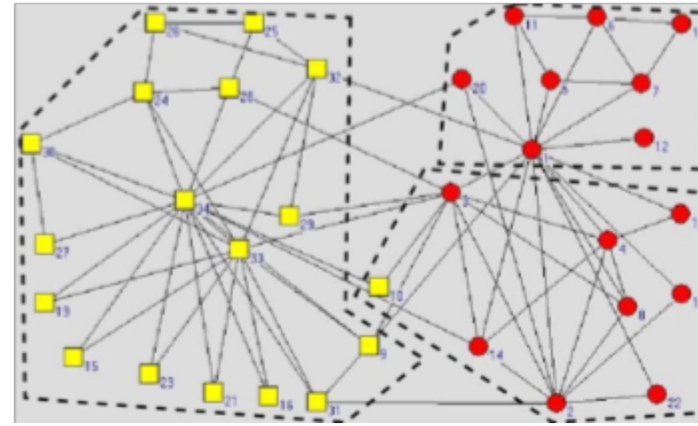
Comparing Initialization Methods

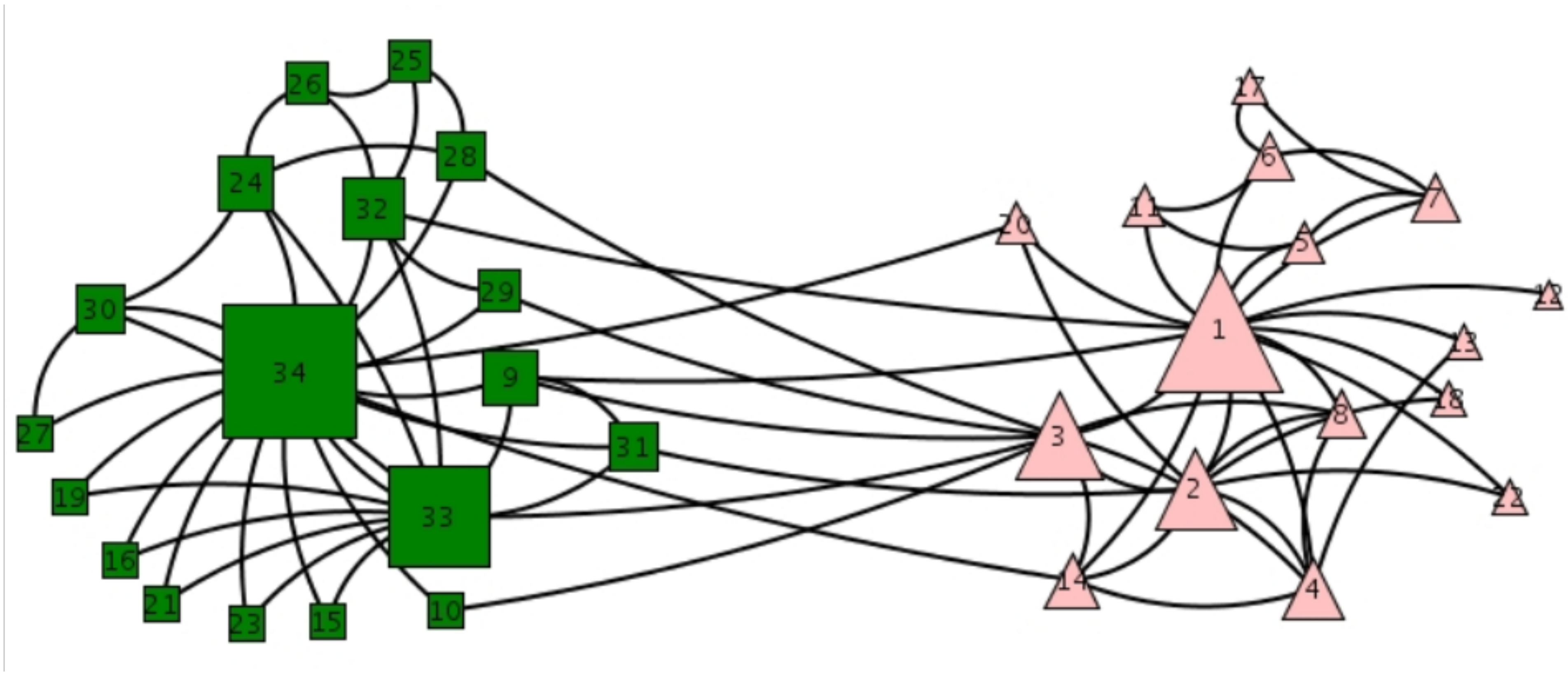
- Naïve
- Top Global Leaders (TGL)
- Top Leaders & not Direct Neighbour (TL&NDN)
- Top Leaders & Few Neighbours in Common (TL&FNiC)

method	dataset	ARI	purity	Q
Naïve	Karate	$.80 \pm .33$	$.90 \pm .20$	$.28 \pm .13$
	Strike	$.59 \pm .25$	$.81 \pm .13$	$.41 \pm .12$
	Football	$.39 \pm .12$	$.66 \pm .08$	$.27 \pm .07$
TGL	Karate	1.0	1.0	0.37
	Strike	1.0	1.0	.54
	Football	.83	.88	.43
TL&NDN	Karate	1.0	1.0	0.37
	Strike	1.0	1.0	.54
	Football	.78	.88	.42
TL&FNiC	Karate	1.0	1.0	0.37
	Strike	1.0	1.0	.54
	Football	.98	.97	.51

Visualized Results

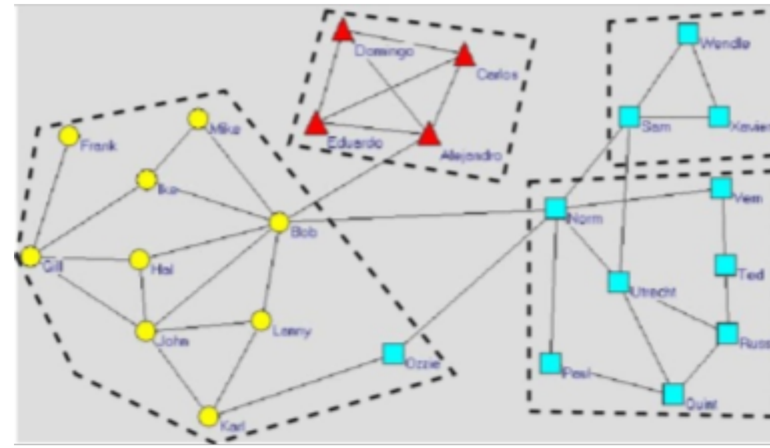
Karate Club

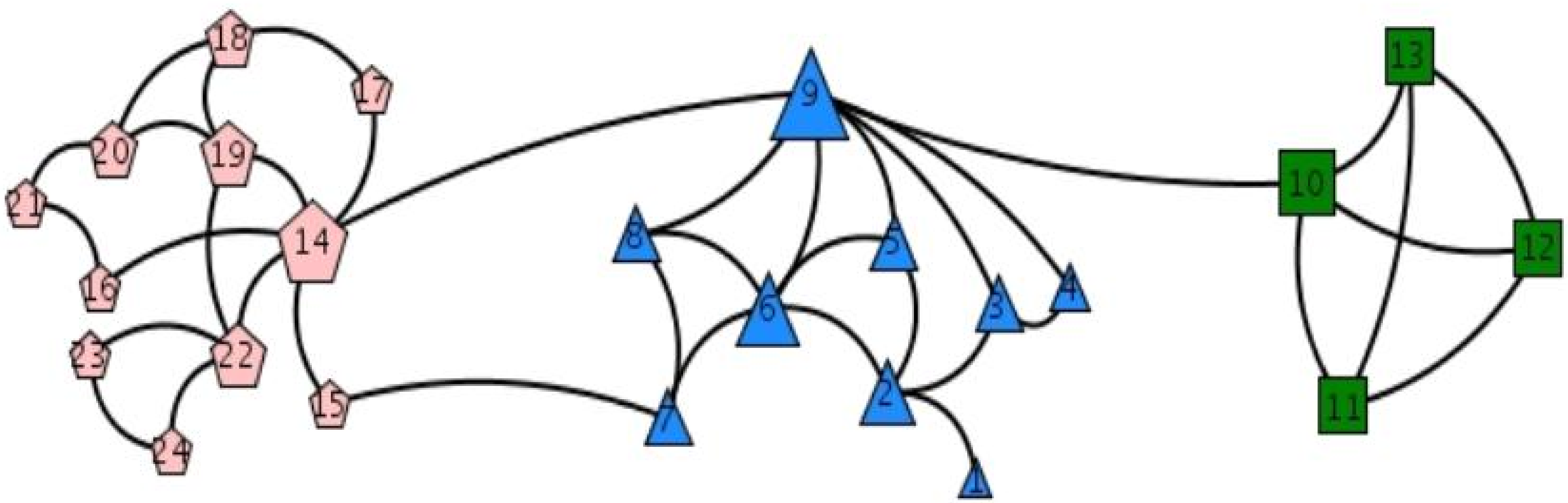




Visualized Results

Strike

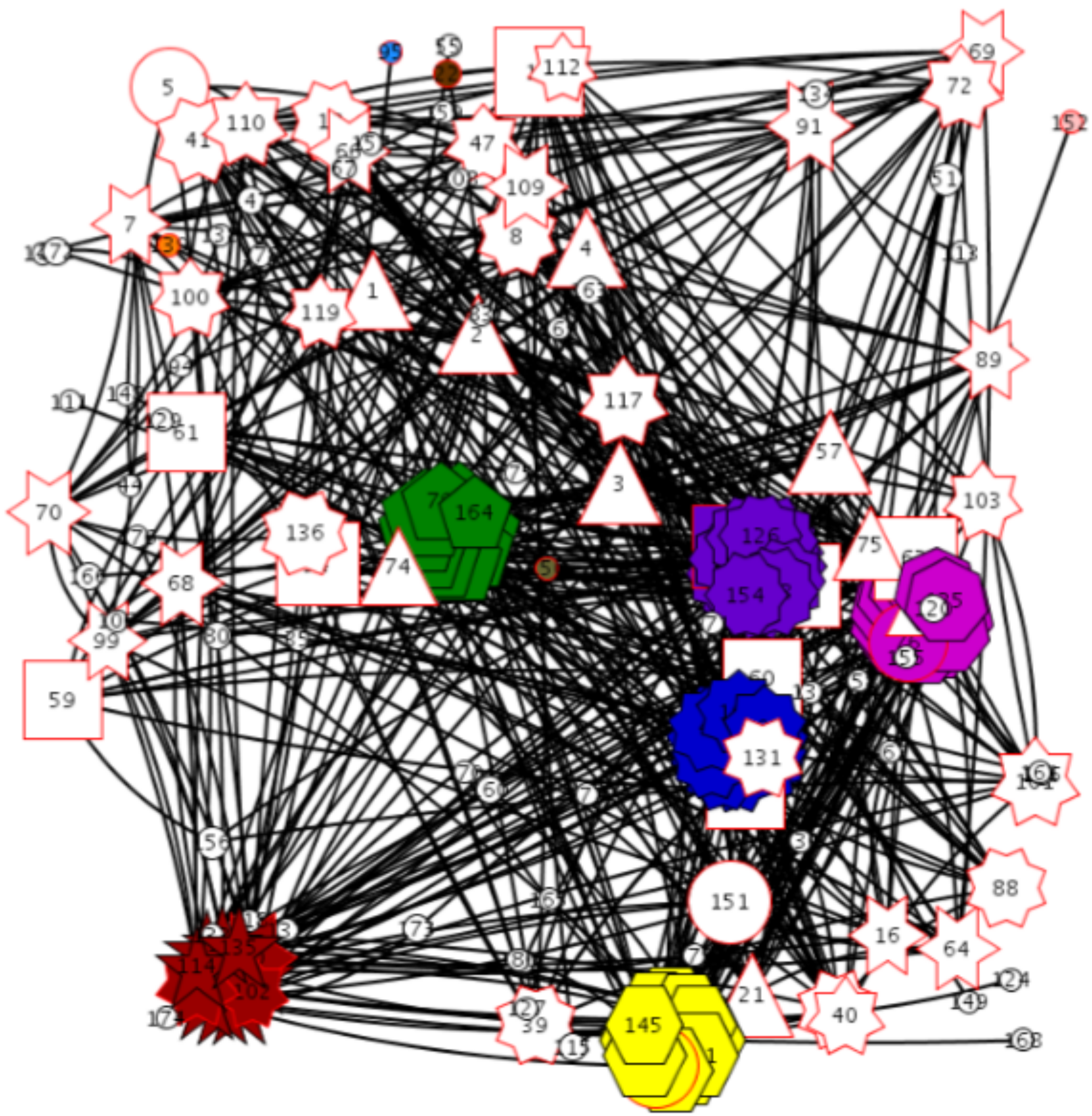




Visualized Results

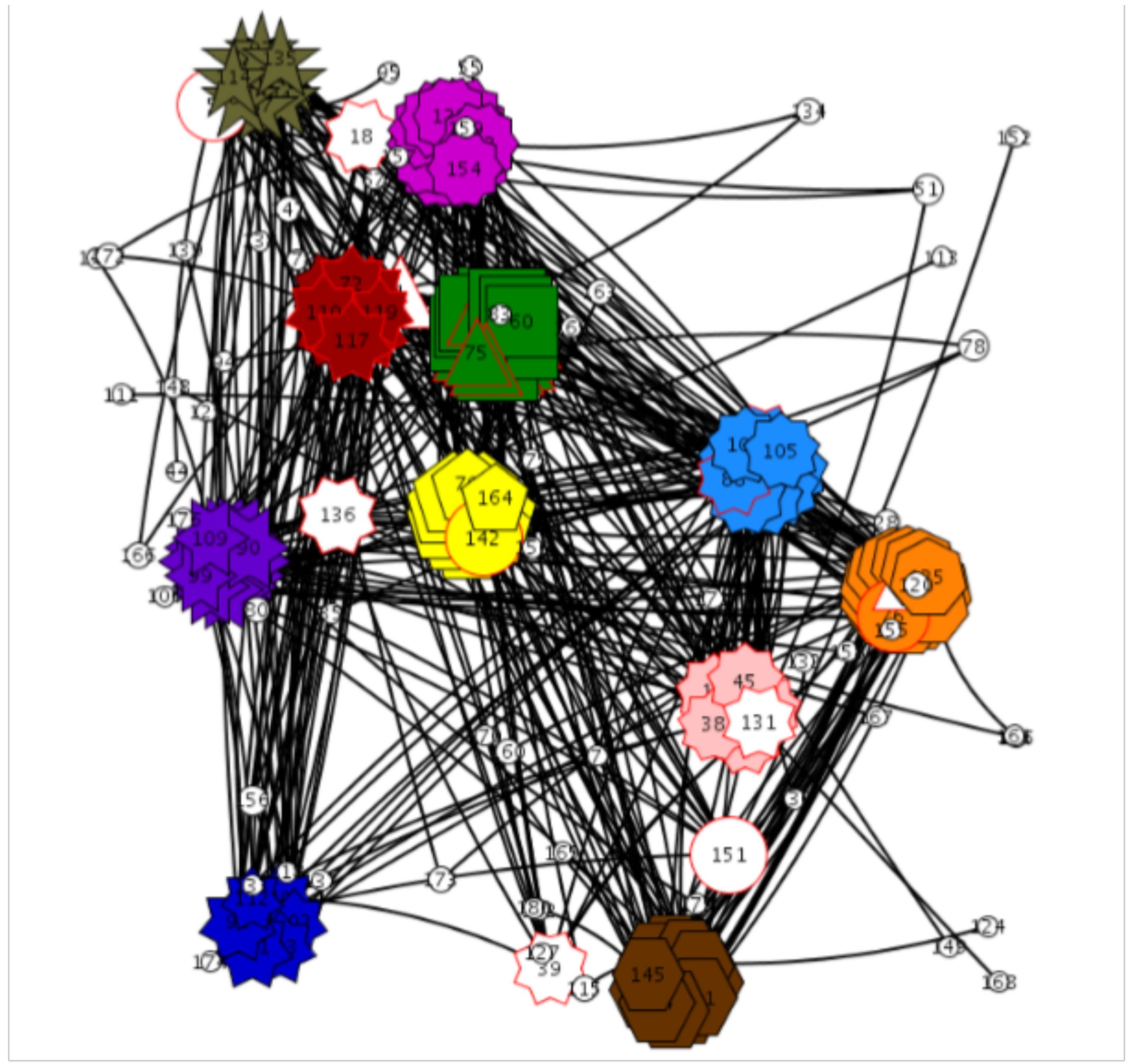
Football, Naïve Initialization

ARI: 0.8±0.3



Visualized Results
Football, Top Global Leaders

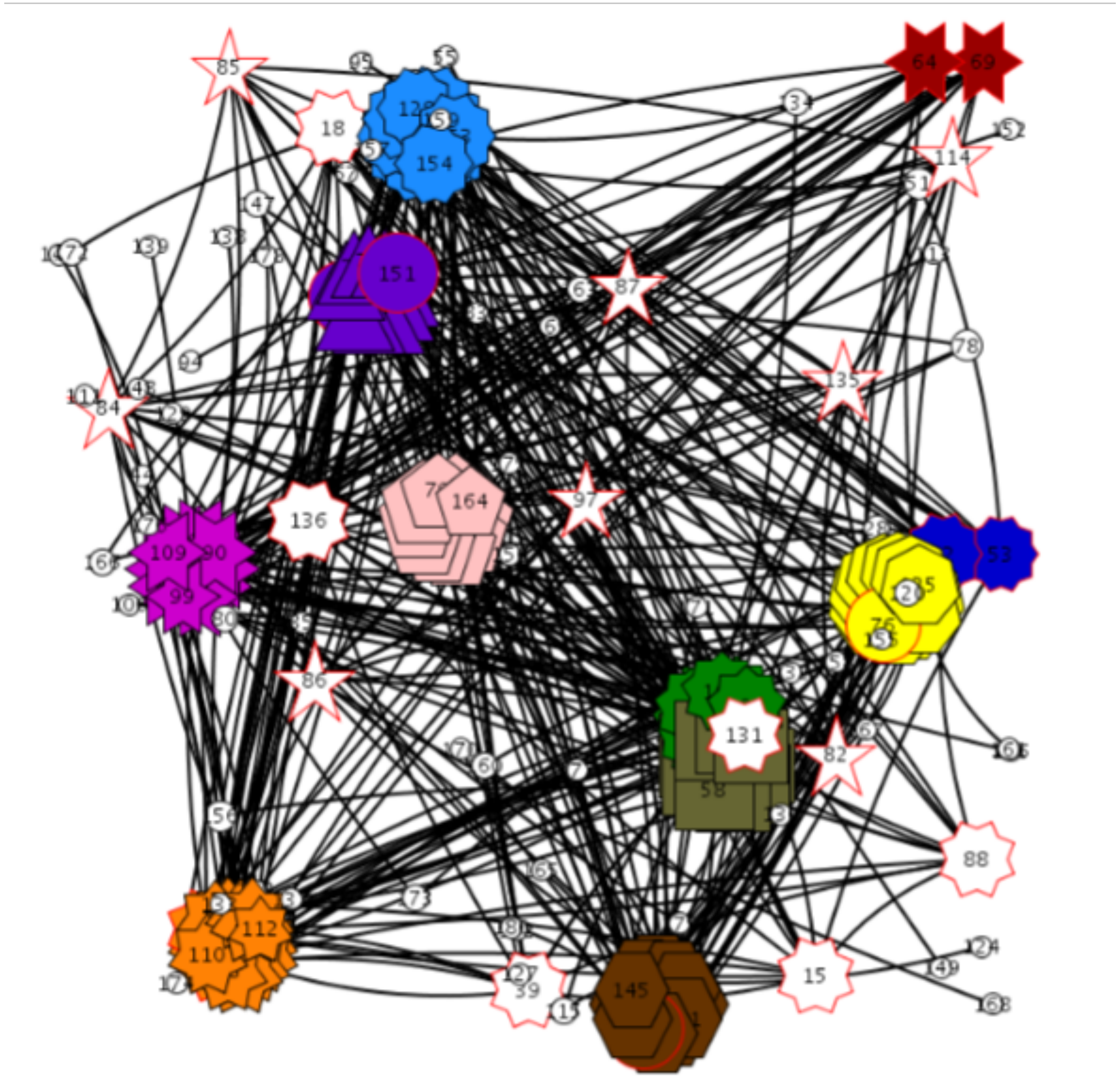
ARI: 0.83



Visualized Results

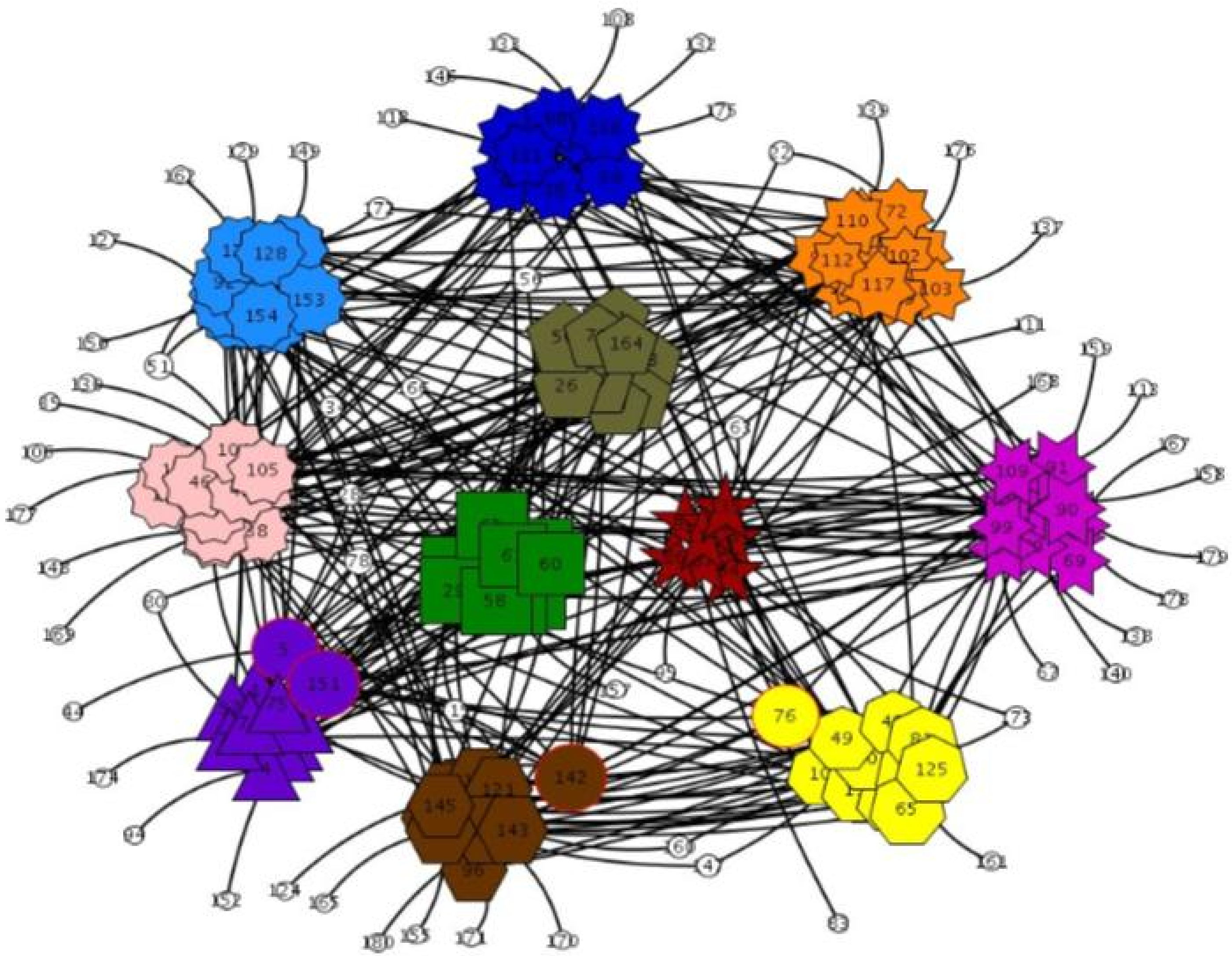
Football, TL & Not Direct Neighbour

ARI: 0.78



Visualized Results
Football, TL & Few Neighbour in Common

ARI: 0.98



Comparing with other approaches

- Given the correct initial k, TopLeaders always provides the best result.
- The other methods do not always find the correct k but even when seeded to Top Leaders, our approach improved the quality of the found communities based on ARI.

dataset	method	k	ARI	purity	Q
Karate 2 groups	fastModularity	3	.680	.970	.380
	cFinder	3	.705	.065	.182
	TopLeader(3)		**.838**	1.0	.374
	SCAN	4	.314	.764	.312
	TopLeader(4)		**.788**	1.0	.361
	TopLeader(2)		1.0	1.0	.371
Strike 3 groups	fastModularity	4	.664	.958	.555
	TopLeader(4)		**.935**	1.0	.532
	cFinder	6	.348	1.0	.485
	TopLeader(6)		**.609**	1.0	.457
	SCAN	3	.848	.958	.547
	TopLeader(3)		**1.0**	1.0	0.548
Football 11 groups	fastModularity	7	.206	.427	.567
	TopLeader(7)		**.637**	.783	.394
	cFinder	12	.983	.913	.532
	TopLeader(12)		**.993**	.977	.511
	SCAN	11	1.0	1.0	.501
	TopLeader(11)		.988	.977	.513

Conclusion

- A novel algorithm to mine communities, which assigns nodes to leaders of communities and selects the leaders of communities iteratively.
- Effective in discovering communities and also in identifying outliers in a network.
- Requires k, the number of desired communities as input. However, it is possible to obtain k after running other contenders and provide the number of discovered communities to our algorithm;
    - Our experimental results showed that communities obtained in this way are more accurate than the original discovered communities even if the used method detected wrong number of communities.

- **Questions?**