Detecting Geographically Dispersed Overlay Communities Using Community Networks

Community structure in networks

- Community is a group of nodes that have a higher likelihood of connecting to each other than to nodes from other communities

- Has numerous applications in social networks and biological networks

- In social networks, understanding the underlying community structure may help to understand the social dynamics in a social group and make predictions (e.g. Zachary's Karate Club **[1]**)

- In biological networks, community structure may help to understand diseases at the cellular level and how groups of molecules carry out cellular functions

Community detection techniques

- **Minimum-cut method** – The network is divided to predetermined number of groups of approximately same size, chosen such that the number of edges between groups are minimized
- **Hierarchical Clustering** – A similarity measure (e.g. cosine similarity) is used to quantify similarity between node pairs
- **Girman-Newman algorithm** – Identifies the edges that connect communities and removes them, based on the betweenness centrality of each edge

Modularity maximization

- Ideal for large, unstructured and self-organizing networks
- Modularity is a scale value between -1 and + 1, which measures the density of edges inside communities to edges outside communities
- Smaller communities are grouped into single nodes and iteratively grown into larger communities until the modularity measure is maximized
- **Louvain method [2]** produces better modularity values with higher performance

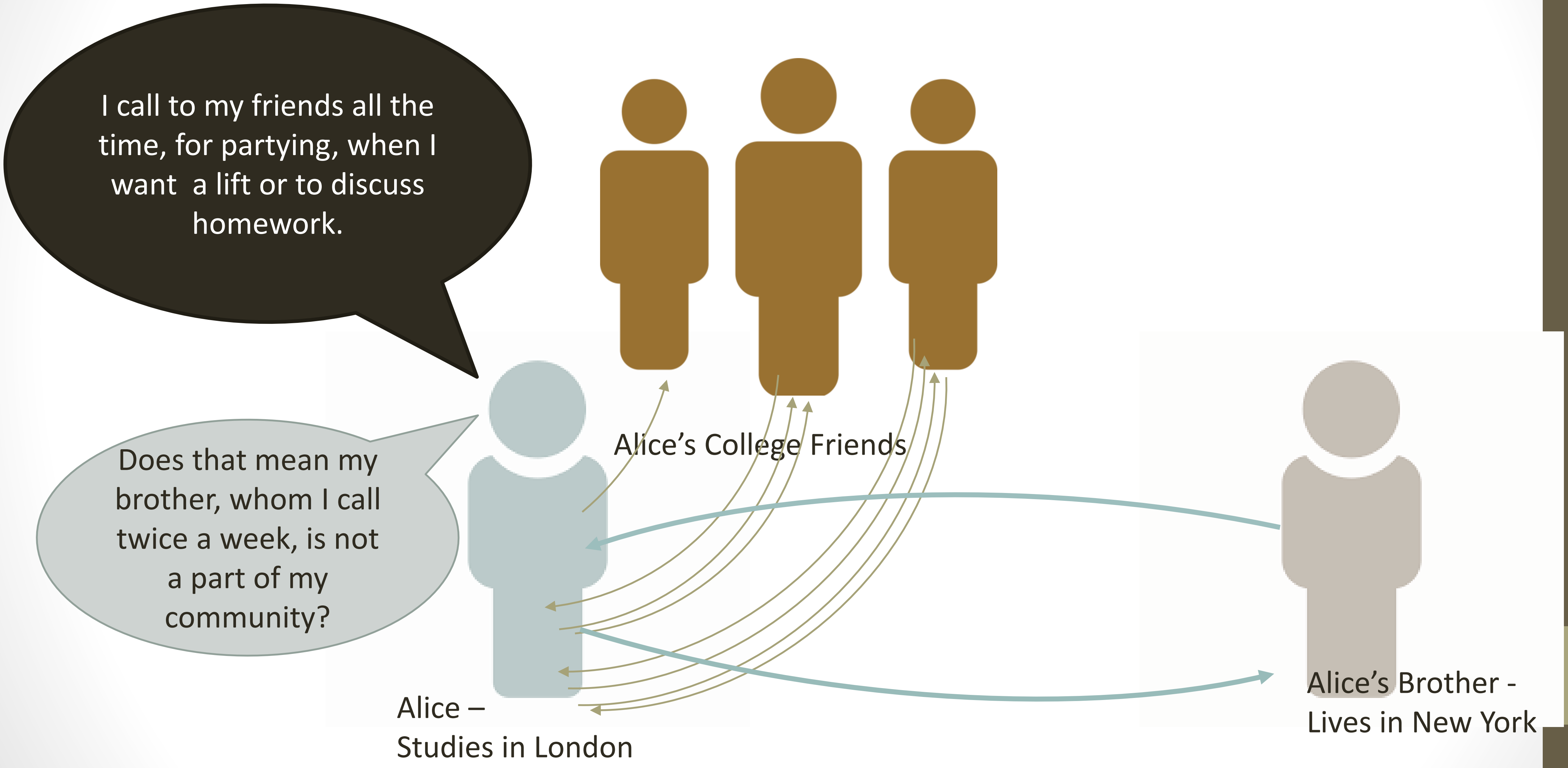
Newman-Girvan Modularity Measure

- The fraction of edges within communities in the observed network minus the expected value of that fraction in a null model

- $M(C) = \frac{1}{2m} \sum_{c \in C} \sum_{i,j \in c} \omega_{ij} - P_{i,j}$ $\qquad P_{ij} = \frac{k_i . k_j}{2m}$

- C is a given partition
- G = (V;E) where V is a set of nodes and E is a set of edges among nodes
- n and m represent the cardinalities of V and E respectively.
- ω_{ij} is the weight associated to each edge (v_j ; v_i)
- For a given node $v_i \in V$, $\eta_i = \{v_j | (v_i ; v_j) \in E \vee (v_j ; v_i) \in E \}$ and $k_i = | \eta_i |$.
- P_{ij} refers to the null model that is used as a reference model, where the edges of the network are rewired randomly while preserving the degree distribution.

Spatial bias in community structure



I call to my friends all the time, for partying, when I want a lift or to discuss homework.

Does that mean my brother, whom I call twice a week, is not a part of my community?

Alice's College Friends

Alice –
Studies in London

Alice's Brother -
Lives in New York

Dist-Modularity [3]

- In many real world social/biological networks the nodes that are in close geographical proximity have a higher tendency of forming communities.
- **Dist-modularity** tries to normalize the effect of spatial bias

$$M_{dist}(C) = \frac{1}{2m}\sum_{c \in C}\sum_{i,j \in c} \omega_{ij} - P_{i,j}$$

$$P_{ij} = \frac{\hat{P}_{ij}. + \widehat{P_{ji}}}{2}$$

$$\widehat{P_{ij}} = \frac{k_i k_j f(d(v_i, v_i))}{\sum_{v_{q \in V}} k_q f(d(v_q, v_i))}; f: R^+ \rightarrow (0,1]$$

- *f*: distance decaying function
- *d*: distance between the two nodes connected by an edge

Detecting geographically dispersed communities

- There can be important links between communities that are geographically far apart.

  E.g: Migrant worker community networks and Terrorist networks

- Dist-modularity tries to normalize the effect of geographical proximity to extract geographically dispersed communities

- However, this is done at the expense of losing the information about the geographically proximate communities. (assumes the community is dispersed at the node level)

- Communities may be geographically dispersed at the community level and not the individual node level

Can we extract geographically dispersed communities **while preserving** the information about the geographically proximate communities???

Extracting geographically distributed overlay communities

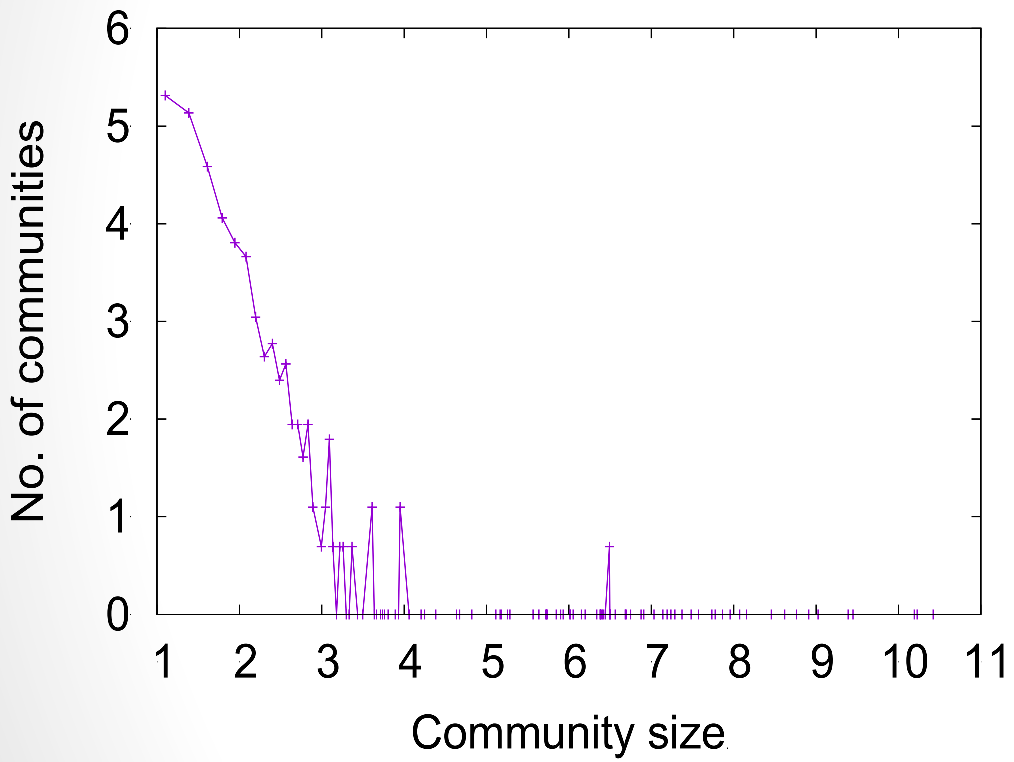
1. Extract the community set C using the Louvain method of N-G modularity optimization;
2. ***for each*** *community c in the set of communities C* **do**
    1. Identify the centroid of each community based on geographical location of each node in the community ;
    2. Assign the centroid as the node representing that particular community in the community network ;
3. ***for each*** *community pair p in the set of communities C* **do**
    1. Compute the strength of the link connecting the community pair p by aggregating the connections among the nodes in community pair p ;
    2. Normalize the link strengths by the community sizes by dividing the link strengths by the multiplication of community sizes of the community pair p ;
4. Identify the communities that are relatively further apart geographically yet have relatively higher link strengths as the **`overlay communities'** ;

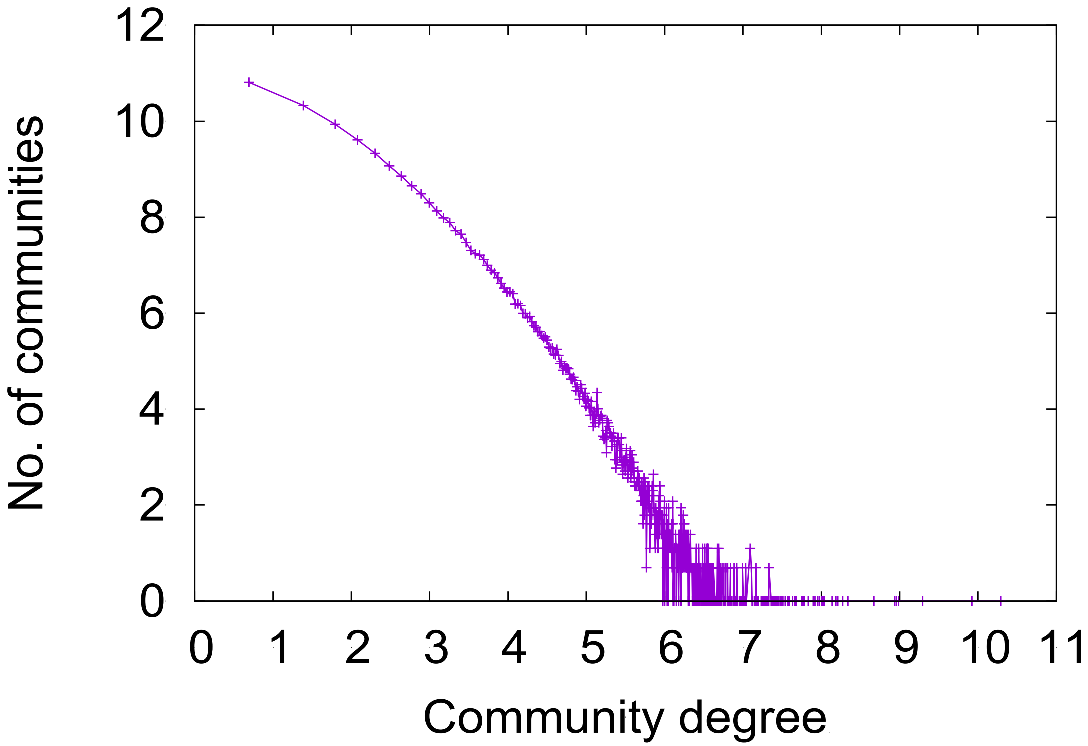
Applying to a real-world network

- Gowala Social Network **[4]** with check-in details of users
- 196,591 connected users with location records of 107,092 users
- Louvain algorithm was applied to total network data set
    - 820 communities were detected at the highest modularity value
- The centroid of each resulting community was decided by home locations of members with known locations (through Check-in details)

Distribution of Community Size and Degree

Evidence for scale-free characteristics of the network

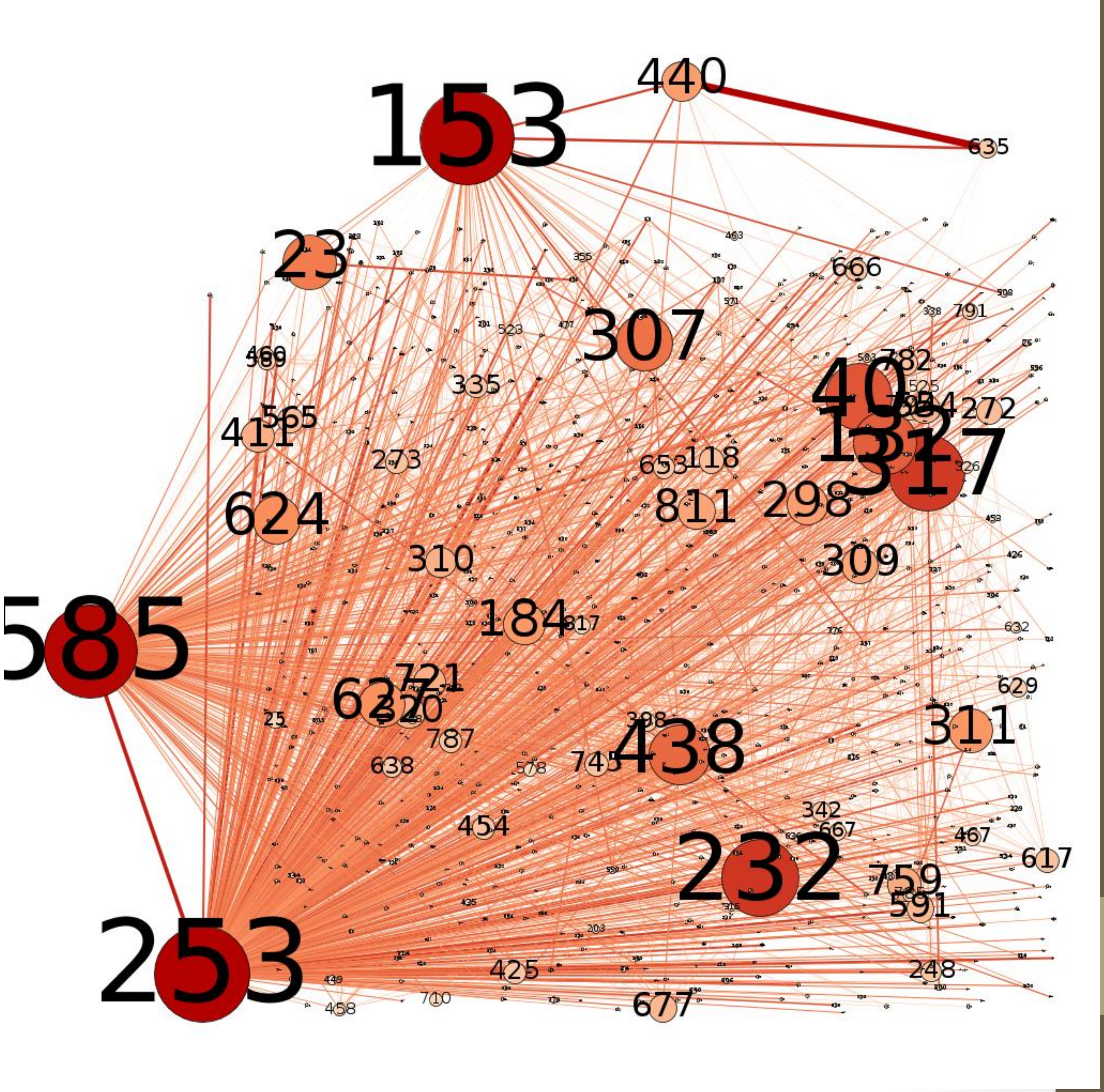




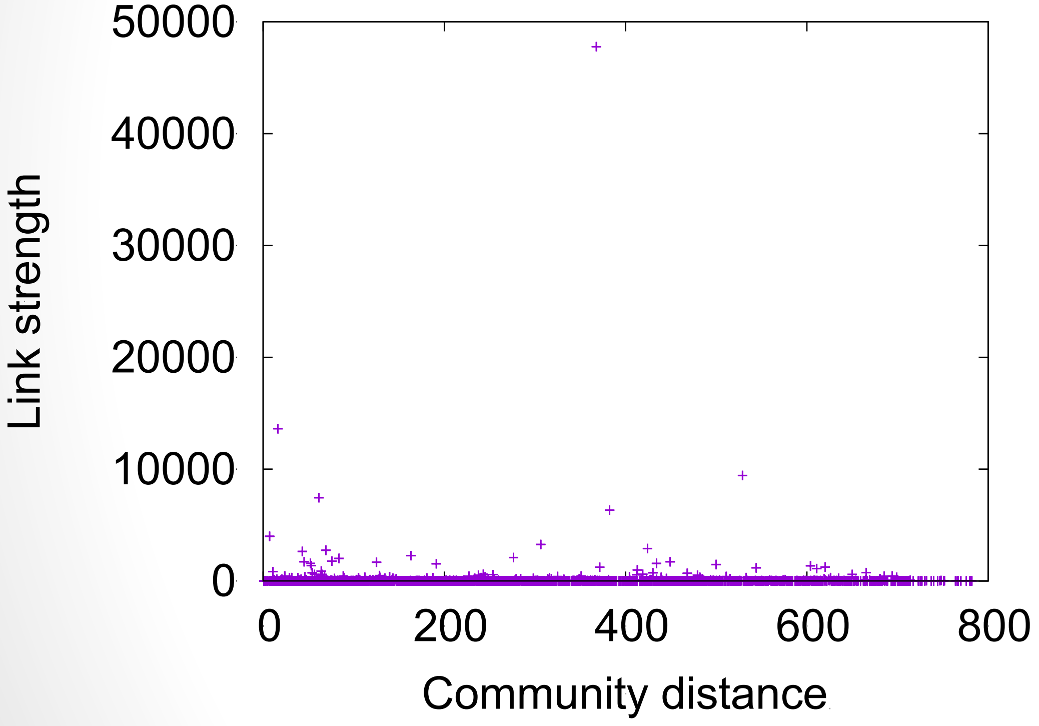
Scale-free correlation: 0.74
Scale-free exponent : 0.67

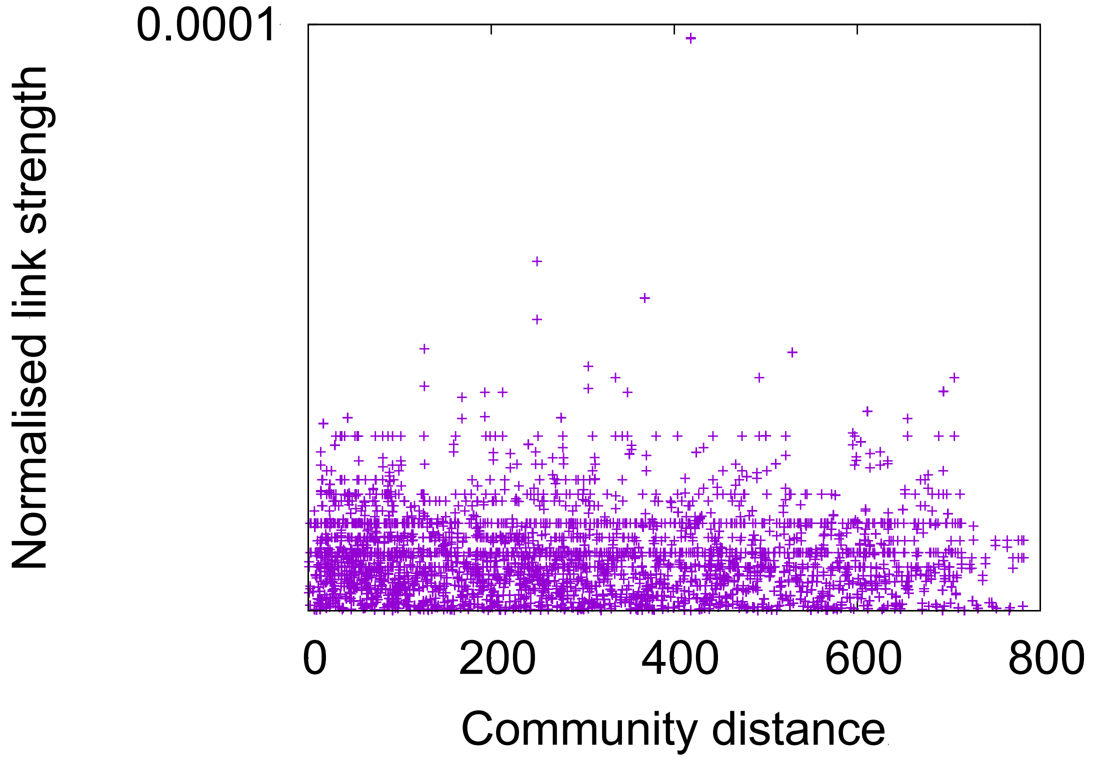
Community network with heterogeneous node sizes and normalized link strengths.

The community sizes and link strengths are non-correlated



Evidence of communities that are tightly connected despite being geographically apart.





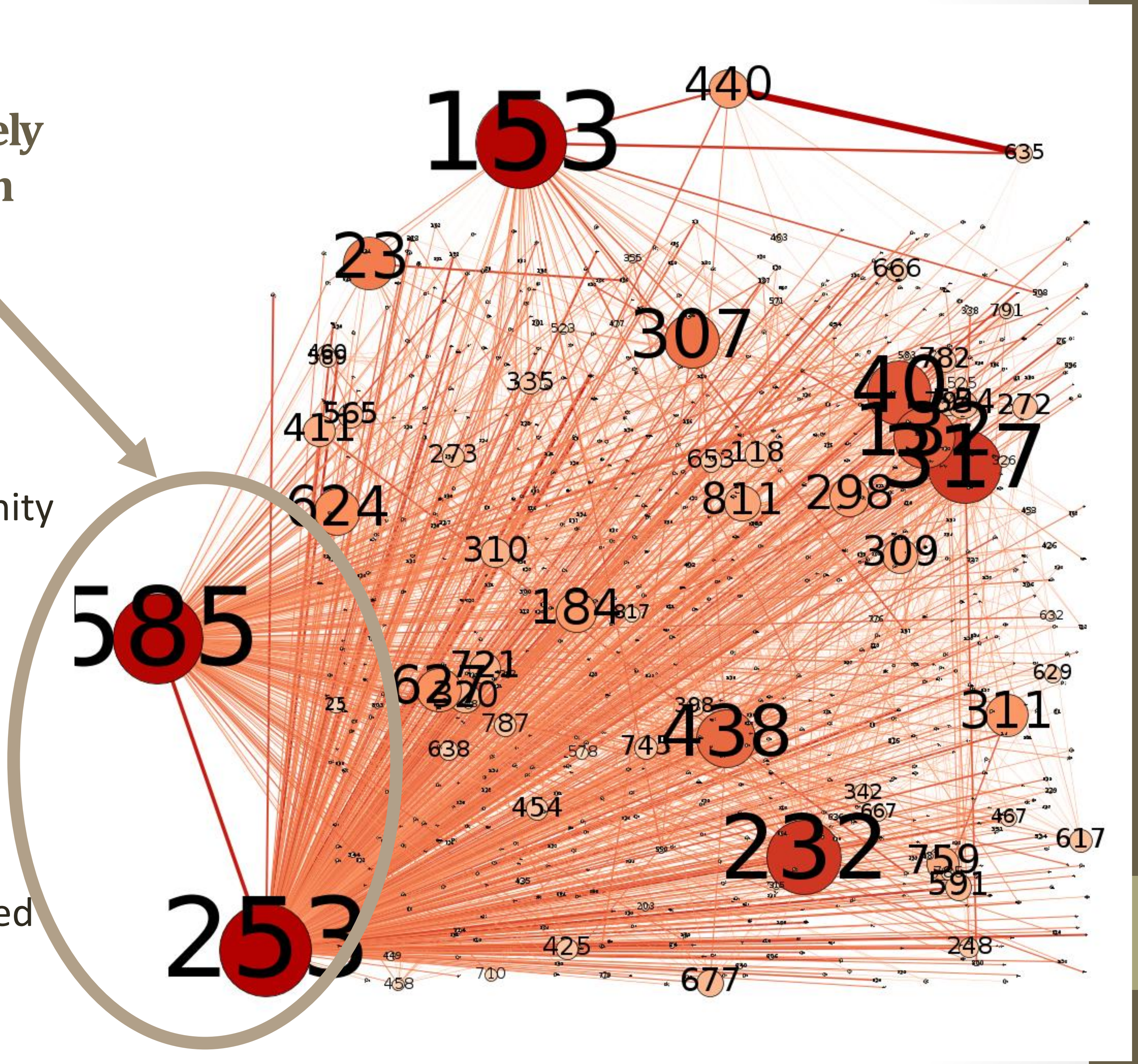
Community Pair with a relatively high normalized link strength

253 - 585

When we normalize the tie strength by population, the link strength of these communities is higher than **82%** of community pairs observed.

It is important to note that these two communities are not overlapping and geographically apart

Thus, we could identify this particular community pair as a geographically dispersed single overlay community.



Extracting geographically distributed overlay communities

- Better performance than dist-modularity optimization –
  Time Complexity **O(n log n)**

- May be used to identify geographically dispersed communities while preserving geographically proximate communities (both may be relevant)

  E.g. Migrant communities, Terrorist cell networks

- The extracted overlay communities may be used for effective marketing campaigns, defense related applications, understanding how economies work, studying migration patterns, etc.

Future work

- Could be applicable in biological networks as well (e.g. Neural networks in the brain)

- More insights from overlay community networks? (centrality, robustness, assortativity)

- Other dimensions for community bias that may be considered other than spatial proximity? (e.g. income/educational level forming a bias in community structure)

Acknowledgements

- LIRNEasia.net
- International Development Research Centre (IDRC) of Canada
- Sri Lanka Institute of Information Technology (www.sliit.lk)

References

[1] Girvan, Michelle, and Mark EJ Newman. "Community structure in social and biological networks." *Proceedings of the national academy of sciences* 99.12 (2002): 7821-7826.

[2] De Meo, Pasquale, et al. "Generalized louvain method for community detection in large networks." *Intelligent Systems Design and Applications (ISDA), 2011 11th International Conference on*. IEEE, 2011.

[3] Shakarian, Paulo, et al. "Mining for geographically disperse communities in social networks by leveraging distance modularity." *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2013.

[4] Leskovec, Jure, and Andrej Krevl. "{SNAP Datasets}:{Stanford} Large Network Dataset Collection." (2015).

why
? wha
hen ?
when
who ? where
what
why
? QUESTIONS
who
?
how
where
who
?
what
how
how
why
? when
where

THANK
YOU!