Assessing Network Partitioning

David G. Robinson
Sandia National Laboratories, Albuquerque, NM

SAND2012-1628C

Overview

Problem

- Complex networks are at the core of much of our world: technological (worldwide web), biological (metabolic pathways), and of course, social networks.
- Identifying clusters or communities within the structure of the network is critical to understanding the functional characteristics of the network.
- Questions in the structural investigation:
  - How well is our model performing at identifying communities?
  - Are there unique or significant clusters? How similar are the communities within a network?
  - Can we measure similarity between communities across networks?

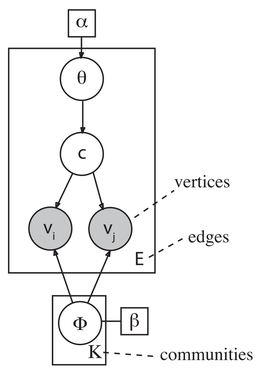
Approach

- Use a hierarchical Bayesian model to identify communities
- Posterior predictive checking to statistically assess the community identification model
- Explore various discrepancy functions for gaining insight into the structure of the network

Community Identification

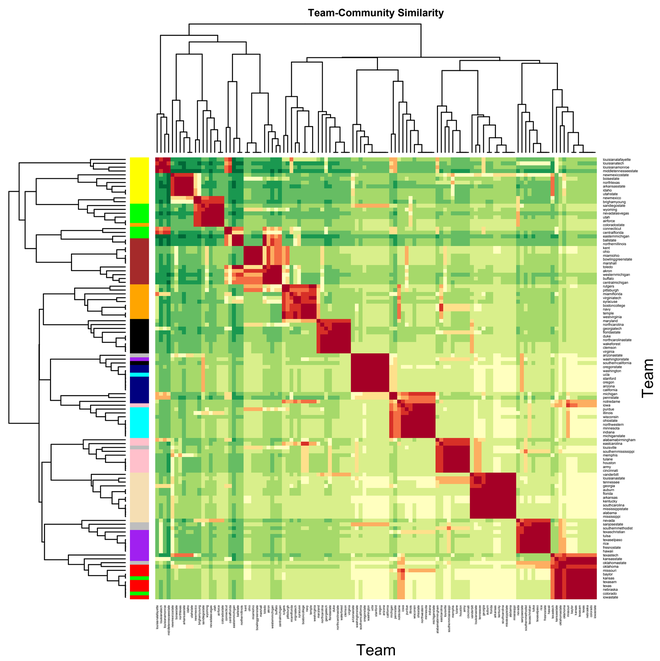
For this initial effort assume the simplest approach and use latent Dirichlet allocation to identify the K communities in a graph G defined by a set of vertices V and a set of edges E [ref]. **For each edge**, $i \in [1, L]$:

- Choose a distribution of communities for each edge, $\Theta = (\theta_1, \ldots, \theta_k) \sim Dirichlet(\alpha)$
- For each of the N vertices
  - choose the community distribution k for each vertice: $z_k \sim Multinomial(\Theta; 1)$
  - choose vertice $v_j \sim p(v_j|z_k, \beta)$ where $\beta_{jk} = p(v_j|z_k)$



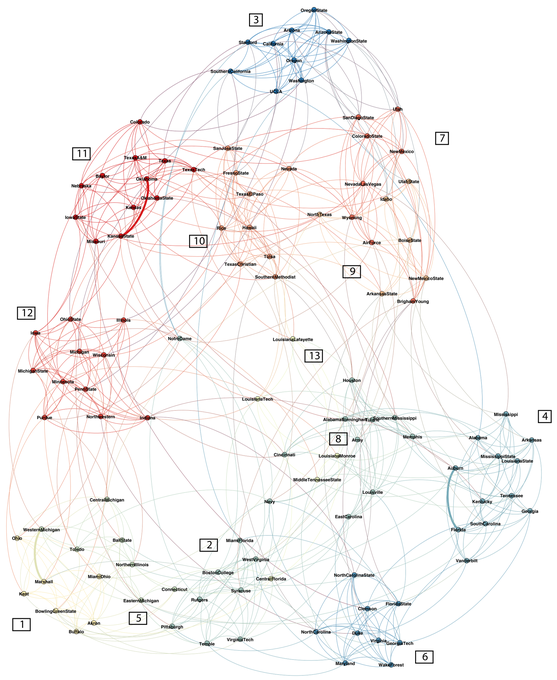
Results: Division IA football games for Fall 2000 [ref]

- 115 vertices (teams), 616 edges (games)
- Identify the communities within the network, then
- Use pair-wise mutual information to understand similarities between the network communities, and
- Cluster similar communities together



Pair-wise Mutual Information for Vertices

Resulting Network



Top 10 Vertices in Each Community

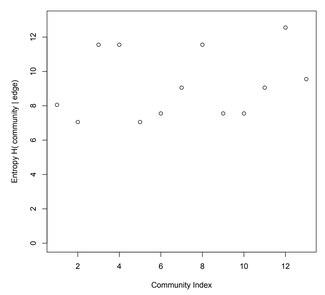
community 1 (6)	community 2 (1)	**community 3 (8)**	**community 4 (9)**	community 5 (6)	community 6 (0)	community 7 (7)
bowlinggreenstate	temple	*southerncalifornia*	*auburn*	easternmichigan	duke	nevadalasvegas
marshall	westvirginia	*washington*	*florida*	northernillinois	northcarolinastate	utah
miamiohio	virginiatech	*arizona*	*southcarolina*	ballstate	wakeforest	airforce
ohio	bostoncollege	*oregon*	*alabama*	toledo	georgiatech	coloradostate
kent	syracuse	*california*	*mississippi*	westernmichigan	clemson	brighamyoung
akron	navy	*stanford*	*mississippistate*	centralmichigan	floridastate	wyoming
buffalo	miamiflorida	*ucla*	*arkansas*	connecticut	virginia	sandiegostate
centralmichigan	rutgers	*oregonstate*	*kentucky*	buffalo	northcarolina	newmexico
pittsburgh	pittsburgh	*washingtonstate*	*tennessee*	centralflorida	maryland	nevada
westernmichigan	notredame	*arizonastate*	*vanderbilt*	akron	vanderbilt	illinois

community 8 (4)	community 9 (10)	community 10 (11)	*community 11 (3)*	*community 12 (2)*	community 13 (10)
houston	arkansasstate	tulsa	kansasstate	wisconsin	middletennesseestate
army	northtexas	rice	nebraska	michiganstate	louisianamonroe
tulane	utahstate	fresnostate	texas	minnesota	louisianalafayette
cincinnati	idaho	texaselpaso	colorado	northwestern	louisianatech
memphis	boisestate	hawaii	texasam	ohiostate	centralflorida
louisville	newmexicostate	texaschristian	iowastate	indiana	connecticut
southernmississippi	texastech	southernmethodist	kansas	illinois	pennstate
alabamabirmingham	newmexico	nevada	baylor	iowa	alabamabirmingham
eastcarolina	brighamyoung	sanjosestate	oklahoma	purdue	michigan
navy	oklahoma	louisianatech	missouri	michigan	miamiflorida

We have intentionally increased the number of assumed communities (13) relative to the true number (12) to corrupt the very discrete underlying community structure. Red highlights indicate vertex not in same conference as other vertices in community.

Community Entropy

The community-edge entropy $H(k|e)$ characterizes how edges are shared or hoarded by communities. The fewer edges shared between communities relative to the number of edges inside the community, the higher $H(k|e)$.



However: the entropy $H(k|e)$ is confounded by possible dependency between edges and vertices; independence is evaluated using a discrepancy function.

Bayesian Model Checking

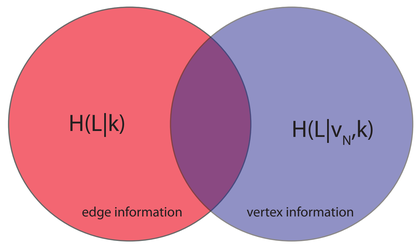
Posterior predictive checking can be used to assess the validity of a Bayesian model without specifying an alternative model [Gelman et al 1996].

A *discrepancy function based on mutual information* is used to compare the value of the distribution function presented by the data and the distribution implied by the posterior.

- Check agreement between observed data and assumed model,
- Identify what portions of the posterior model better fit the observed data and
- provide insight into where community structure should be explored further.

Discrepancy Function

We can characterize the independence of edges and vertices through a discrepancy function that measures the instantaneous mutual information between the vertices V and edges E:



$$\mu_I(v_N, E|k) = H(L|k) - H(L|v_N, k)$$

Discussion

The first term in μ_i, $H(L|k)$ is the edge entropy:

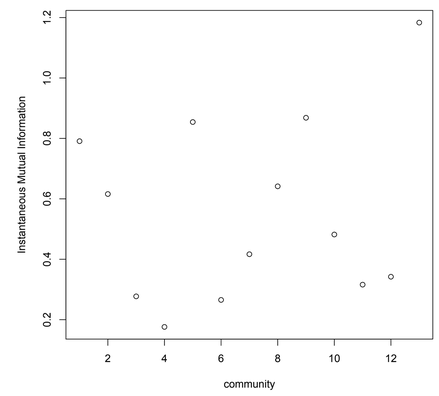
- The more edges that are associated external to a community, the higher the entropy $H(L|k)$, but
- If edges are limited to a few communities then the edge entropy is low

The second term is related to the entropy of the vertices:

- If a vertice is associated with only a few edges, it will have lower entropy over the edges, resulting in a low value of $H(L|v_N, k)$ relative to $H(L|k)$ and μ_i will be high.
- If a vertice is associated with many edges, $H(L|v_N, k)$ will be close to $H(L|k)$ and μ_i will be low.

Communities that are significant within the network have low mutual information μ_i between vertices and edges.

Results



- Note the high μ_i for Community 13, indicating the lack of 'uniqueness' of this community;
- Communities 3,4, and 6 have the lowest μ_i indicating the communities are statistically significant within the network.

Conclusions

- A very general information based approach to determining the statistical significance of communities within a network has been developed and used to investigate a well understood problem as proof of concept.
- Subsequent efforts will involve application to more complex networks and comparison with the algorithm developed by Lancichinetti, et al

References

- Gelman, Meng, Stern, 1996, *Posterior Predictive Assessment of Model Fitness via Realized Discrepancies.*
- Lancichinetti, Radicchi, and Ramasco, 2009, *Statistical Significance of Communities in Networks.*