0.1 Comparing Community Structure with Normalized Mutual Information

In the previous section, we saw that the large scale statistics of the communities were highly dependent on the type of community under consideration. However, macroscale network statistics do not account for differences in community structure that result from operations such as splitting or merging of communities. Moreover, this view does not account for which users belong to which communities, and in particular which users belong to the same communities across community types. To answer this question, we invoke methods for cluster comparisons: given two different clusterings of nodes into communities, how similar are the two clusters? The standard approach to answering this question is to define a metric on the space of clusterings. Because OSLOM detects *coverings* rather than *partitions* of users, standard cluster comparison metrics like variation of information [**?**] are not appropriate. Instead, we use a generalization of variation of information first introduced in [**?**], the normalized mutual information. The normalized mutual information stems from treating clustering as a community identification problem: given that we know a node's community membership(s) in the first clustering, how much information do we have about its community membership(s) in the second clustering, and vice versa? Consider the two coverings $\mathcal{C}_1$ and $\mathcal{C}_2$.We think of the community memberships of a randomly chosen node in $\mathcal{C}_1$ as a binary random vector $\mathbf{X} \in \{0,1\}^{|\mathcal{C}_1|}$ where the i^{th} entry of the vector is 1 if the node belongs to community i and 0 otherwise. Similarly, $\mathbf{Y} \in \{0,1\}^{|\mathcal{C}_2|}$ is a binary random vector indicating the community memberships of the node in $\mathcal{C}_2$. Then the normalized mutual information is defined as

$$\text{NMI}(\mathcal{C}_1, \mathcal{C}_2) = 1 - \frac{1}{2}\left(\frac{H[\mathbf{X}|\mathbf{Y}]}{H[\mathbf{X}]} + \frac{H[\mathbf{Y}|\mathbf{X}]}{H[\mathbf{Y}]}\right) \tag{1}$$

where $H[\cdot]$ is a marginal entropy and $H[\cdot|\cdot]$ is a conditional entropy. The normalized mutual information varies from 0 to 1, attaining the value of 1 only when $\mathcal{C}_1$ and $\mathcal{C}_2$ are identical coverings up to a permutation of their labels. See the appendix of [**?**] for more details.

We considered the normalized mutual information between the communities inferred from the structural network and the networks weighted with lag 1 through 6 transfer entropies, hashtag similarity, and mention, retweet, and mention-retweet activity. The resulting $\text{NMI}(\mathcal{C}_i, \mathcal{C}_j)$ are shown in Figure 1. We see that similarity between the coverings is dictated by the generic community type (structural, activity-based, etc.). That is, the transfer entropy coverings are more similar to each other than to any of the other coverings, with a similar result for the mention, retweet, and mention-retweet coverings. Interestingly, the coverings resulting from the different weightings are all more similar to each other than to the structural covering from the unweighted network. Also note that the covering based on the hashtag similarities are different from all of the other weight-based coverings.

Thus, we see that although the activity-based, interaction-based, and topic-based communities relied on the structural network, their community structure differs from *the most* from the community structure of the follower network. This agrees with the results from the previous section, and reinforces that the follower network is a

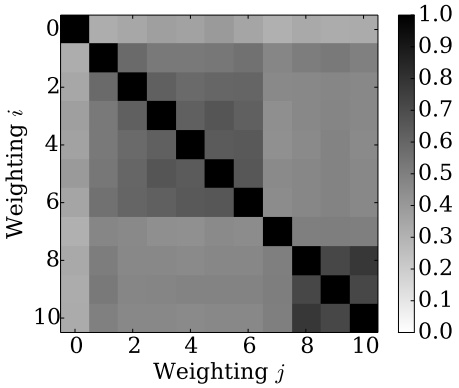

Figure 1: The normalized mutual information between the communities using the different weightings. Weighting 0 corresponds to the structural (binary weighting) network, weightings 1 through 6 correspond to the weighting using the transfer entropies with lag 1 through 6, weighting 7 corresponds to the hashtag similarity, and weightings 8, 9, and 10 correspond to the mention, retweet, and mention-retweet weightings. Values of normalized mutual information close to 1 indicate similarity in the community structure, while values close to 0 indicate dissimilarity. The normalized mutual information is computed with the singletons removed.

necessary but not sufficient part of detecting communities in online social networks.