Correlation between linguistic content and social links in an on-line network

As an original source of linguistic data, **computer-mediated communication** (CMC) offers advantages for language research. Exchanges among on-line users, for instance, provide a record of linguistic interactions at a scale that traditional ethnographic methods cannot approximate. Using quantitative methods that exploit the rich structure of CMC, we analyze data from StackOverflow (a question/answer on-line forum) to support the hypothesis that *users tend to participate in exchanges with other users based on the topic or the content of the discussion* (H1). We will show that H1 has consequences for a social theory of communication, in which actors are motivated by the logic of the "commons", not their immediate self interest.

A StackOverflow **posting** is a thread of question/answer messages, often technical in nature. To narrow down our data, we sample 50 postings related to the topic of *vulnerability*. The postings are hand-annotated following a **TEI XML schema** (Beißwenger et *al*. 2012), and organized in a corpus. The annotations include information about the users (handle, reputation score, etc.) and their messages (text, place in a thread, etc.). Using that information, we employ **Social Network Theory** to analyze the interactions among users. We build a **bi-modal network**, with postings and users as nodes. Edges connect users to the postings they contribute to. From that we project a **coaffiliation network**, in which edges connect postings if the same user contributed to them.

To test H1, we investigate whether postings that share social links are semantically similar. We extract all of the language from each posting, treating each posting as an aggregate document. We process each document using familiar empirical **Natural Language Processing** techniques: lowercase normalization, removal of stopwords (frequent but uninformative words). The resulting 9000+ terms are the dimensions representing each document in a vector space, but we reduce this dimensionality with **Latent Semantic Analysis**. The resulting document-by-term matrix is turned into a document-by-document **distance matrix**, computing the distances between the vectors with a **cosine** function. Applying a **hierarchical agglomerative classifier** to the distance matrix results in 5 distinct semantic clusters , used as node attributes in the coaffiliation network (nodes colored according to cluster in figure 1).

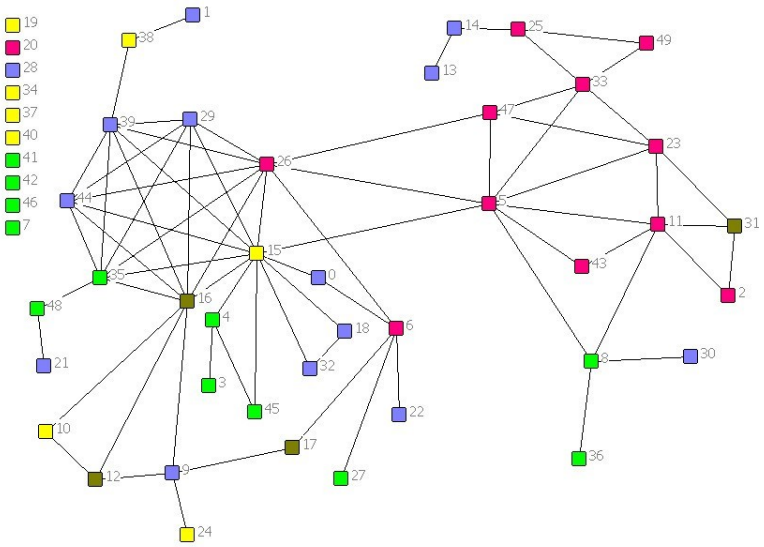

FIgure 1

We then test the degree of **homophily** in the network, fitting an **ANOVA density model** (using the UCINET package). The resulting *p-value = 0.009* shows a significant correlation between semantic class and social link. That is, *users (links) have a tendency (beyond chance) to go across postings (nodes) with similar content*. We take this to be evidence in favor of H1. Our conclusion is that StackOverflow users are not acting opportunistically, but rather make contributions according to their expertise, knowledge, or specific interests. We then argue that the model of CMC that best explains this behavior is that contributions to online communities are not motivated by short-term individual rewards, but by long-term collective gains that are shared by the members of a community (Wasko & Faraj 2005, von Krogh *et al.* 2012).

References:

Beißwenger, M, Ermakova, M, Geyken, A, Lemnitzer, L, & Storrer, A (2012). A TEI Schema for the Representation of Computer-mediated Communication. *Journal of the Text Encoding Initiative* 3. <http://jtei.revues.org/476>; DOI: 10.4000/jtei.476.

von Krogh, G.; Haefliger, S.; Spaeth, S.; & Wallin, M. W. (2012). Carrots and Rainbows: Motivation and Social Practice in Open Source Software Development. *MIS Quarterly* 36(2): 649-676.

Wasko, M. M.; & Faraj, S (2005). Why should I share? Examining social capital and knowledge contribution in electronic networks of practice. *MIS Quarterly* 29(1): 35-57.