Complex networks, semantic links, and affiliations in StackOverflow

R. Duke*, V. Filkov, P. Devanbu, J. D'Souza, and R. Aranovich* (UC Davis)

[*Corresponding authors: {rebrim; raranovich}@ucdavis.edu]

Research project

Questions

- Does the content of communication in online communities predict the development of interconnected social groups?
- Does the existence of an interconnected social group in an online community provide a "true" measure of semantic categories among messages?

Approach

- Build a corpus of StackOverflow posts
- Analyze the relations between posts and users with social network theory

Larger issue: role of linguistic cues in the formation and behavior of developer communities in Open Source software systems.

Background: CMC and online forums

Linguistic features of computer-mediated communication (CMC)

- Hyperpersonal (Walther, 1996)
- Hybrid oral-written (Herring & Paolillo, 2006, Tagliamonte & Dennis, 2008)
- Frequent neologisms
- Parse resistant (Gimpel et al, 2011, Derczynski et al, 2013)

Motivations for users to contribute to forums:

- Ego-centric, or
- Group centric (von Krogh et al, 2012)

Modeling CMC interactions with social network analysis

- Applications of SNA to the study of CMC (Postmes et al. 2000, Haythornthwaite 2005, Wasko & Faraj 2005, Brown et al. 2007)
- The link prediction problem: Use network features to predict where a new link between nodes will develop
- Detection of stylistic and semantic clusters (Postmes et al. 2000)

StackOverflow Corpus

Structure of SOF

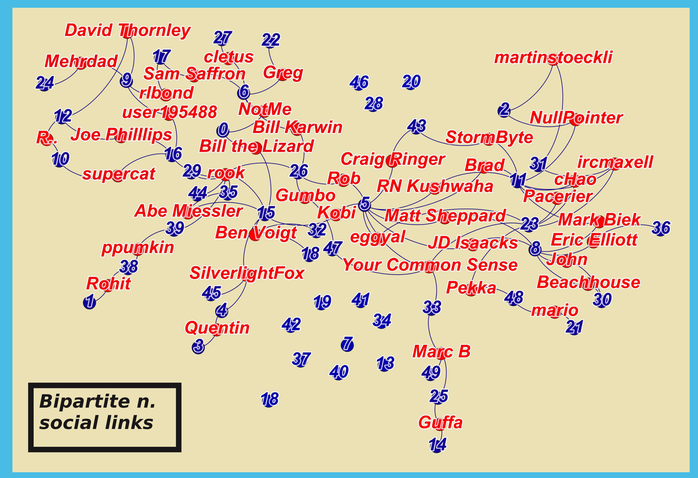
- A user posts a question ("forum"), starting a thread of answers ("answer").
- Discussion is moderated: posts can be messages from moderators ("moderator") to the question
- Users can post comments on "forum", "answer", or "moderator, in the form of a "response".
- Questions are technical in nature, related to computer technology.
- Posts mix natural language (English) with code.
- Users and their posts are ranked by reputation and up/down votes.

Structure of the corpus

- Posts are selected around the term vulnerability
- Size of corpus: 50 Posts, 835 Users, 217,000 Words
- XML-tagged files, based on the CMC TEI schema (Beißwenger et al, 2012)

Network modeling using the XML features

- Exploit information in XML tags to build bipartite networks
- Posts are group actors, linked by roving members (first bipartite network here, social and posts)



Experiment

Hypothesis: social links among posts are more likely to develop if the posts are semantically related.

- Social links measured by users "roving" from one post to another (i.e. making a contribution to more than one post).
- Semantic relations measured by shared keywords

Experimental design

- Generate a separate bipartite network of keywords and Posts (PXK).
- Generate two parallel co-affiliation networks, sharing the same nodes (posts): one for social ties, another one for semantic ties
- Test to see if there is a correlation between the ties in one co-affiliation network and the ties in the other co-affiliation network.
- Statistical test procedure: QAP correlation

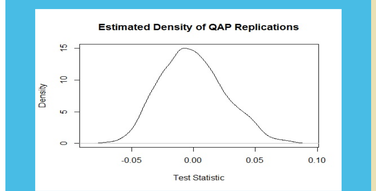
QAP Test Results

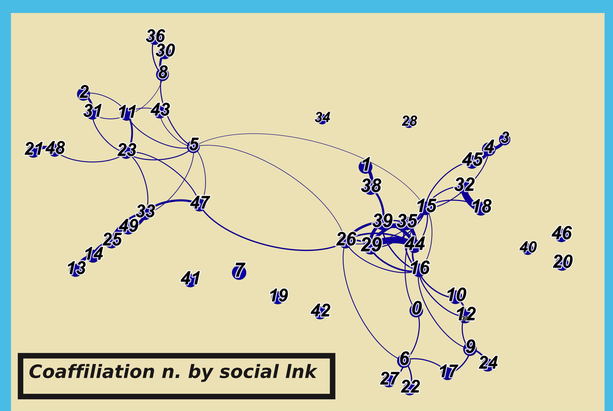
Estimated p-values:

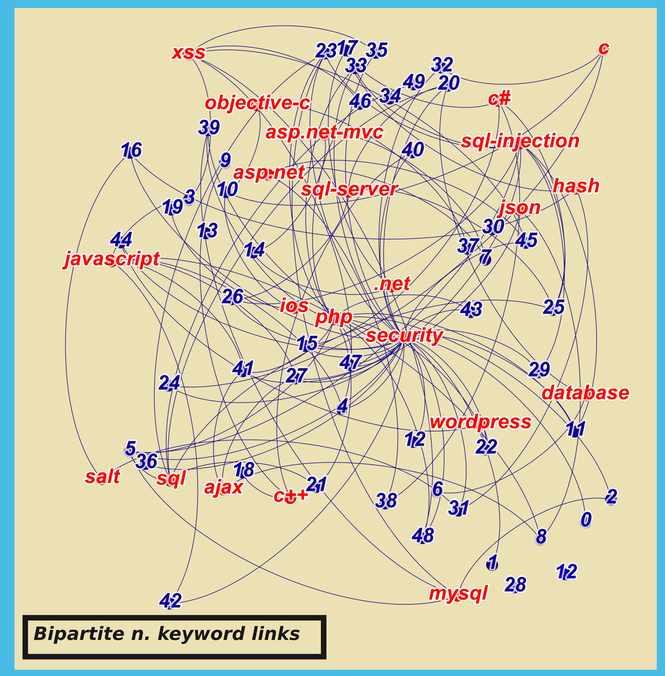
p(f(perm) >= f(d)): 0
p(f(perm) <= f(d)): 1

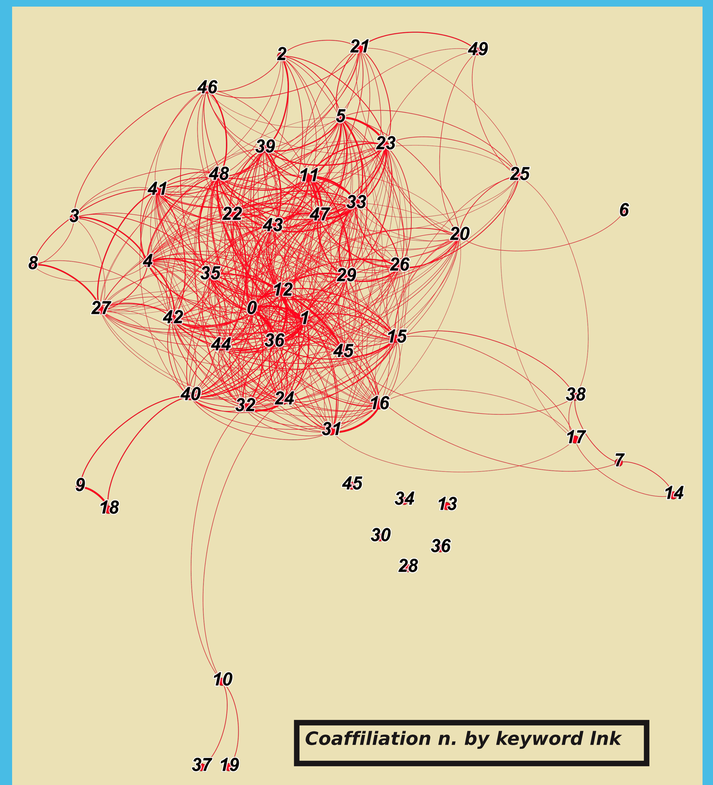
Test Diagnostics:

Test Value (f(d)): 0.211721
Replications: 1000
Distribution Summary:
Min: -0.06862837
1stQ: -0.01765577
Med: -0.004912615
Mean: -0.0005735719
3rdQ: 0.01420211
Max: 0.07791787









Results:

- SOF posts coaffiliation networks by social ties(users) and by keyword ties (semantics) look different from each other
- QAP statistics reveal that the networks are positively correlated
- The correlation is expected under the hypothesis that users rove from post to post following their semantic relations.

Discussion

- Users don't post randomly, staying within topic when roving the network. Users contribute to areas they feel confident in.

- "Consistencies exist in groups interacting via a CMC system, such that variations of content and form of interaction styles will be larger between groups than within groups" (Postmes et al. 2000)

- Keywords provide a mesure of semantic relatedness between documents (i.e. posts) that does not require coding by an external researcher
- Social links can provide a measure of "true" semantic relatedness that can be used to evaluate semantic clustering algorithms and document classification techniques.
- Many "committed" users seem to contribute to the community to increase the "common good", disseminating wisdom and expertise. Such behavior may offer support for the view that contributions to online communitites are not morivated by short-term individual rewards, but by long-term collective gains that are shared by the members of a community (Von Krogh 2010)

REFERENCES

Derczynski, L., Ritter, A., Clark, S., & Bontcheva, K. (2013). Twitter Part-of-Speech Tagging for All: Overcoming Sparse and Noisy Data. In Proceedings of Recent Advances in Natural Language Processing (pp. 198–206). Hissar, Bulgaria.

Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., … Smith, N. A. (2011). Part-of-Speech Tagging for Twitter: Annotation, Features, and Experiments. In Proceedings of the Annual Meeting of the Association for Computational Linguistics, companion volume. Portland, OR.

Haythornthwaite, C. (2005). Social networks and Internet connectivity effects. Information Communication & Society 8(2): 125-147.

Postmes, T.; Spears, R.; & Lea, M. (2000). The formation of group norms in computer-mediated communication. HUMAN COMMUNICATION RESEARCH 26: 341-371.

Tagliamonte, S, Denis, D. (2008). Linguistic ruin? LOL! Instant messaging and teen language. American Speech, 83(1), 3-34.

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

Brown, J.; Broderick, A. J.; & Lee, N. (2007) Word of mouth communication within online communities: Conceptualizing the online social network. JOURNAL OF INTERACTIVE MARKETING 21(3): 2-20.

Walther, J. B. (1996). Computer-Mediated Communication: Impersonal, Interpersonal, and Hyperpersonal Interaction. Communication Research, 23(1), 3–43.

Herring, S. C., & Paolillo, J. C. (2006). Gender and genre variation in weblogs. Journal of Sociolinguistics, 10(4), 439-459.

von Krogh, Georg; Haefliger, Stefan; Spaeth, Sebastian; and Wallin, Martin W.. (2012). "Carrots and Rainbows: Motivation and Social Practice in Open Source Software Development," MIS Quarterly, (36: 2) pp.649-676.

Beißwenger, M, Ermakova, M, Geyken, A, Lemnitzer, L, and Storrer, A (2012), « A TEI Schema for the Representation of Computer-mediated Communication », Journal of the Text Encoding Initiative [Online], Issue 3. 2012, Online since 15 October 2012, connection on 16 May 2016. URL : http://jtei.revues.org/476 ; DOI : 10.4000/jtei.476

Wasko, M., Teigland, R., & Faraj, S. (2009). The provision of online public goods: Examining social structure in an electronic network of practice.