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

Raúl Aranovich, Rachael Duke,
Joshua Gomez, Vladimir Filkov.

University of California Davis
Research supported by NSF
Award #1445079

Introduction

- **General question**: Why do busy people choose to be active participants in Public Virtual Communites (PVC)?

a. Simple answer: direct individual benefit

b. Alternative answer (Wasko & Faraj 2005, Von Krogh et al. 2012): indirect reward, by being members of a team that reaps collective benefits .

Introduction

- **Hypothesis 1**: If indirect rewards matter more than direct benefits, users in a PVC should be more likely to contribute to threads that are semantically related.
- **Methods**
  - Build a network of users × threads (documents)
  - Documents are classified according to their content, using NLP techniques.
  - Edges should develop between users and documents of similar semantic class.
- **Expectation**: A network modeling a PVC will have a community structure determined by the content of the documents.

Introduction

- **Focus of investigation**: communities in the online technical forum StackOverflow

- StackOverfow
  - "...a community of 7.2 million programmers, just like you, helping each other."
  - Question-answer threads, member-generated.
  - Up/down answer votes determine reputation.

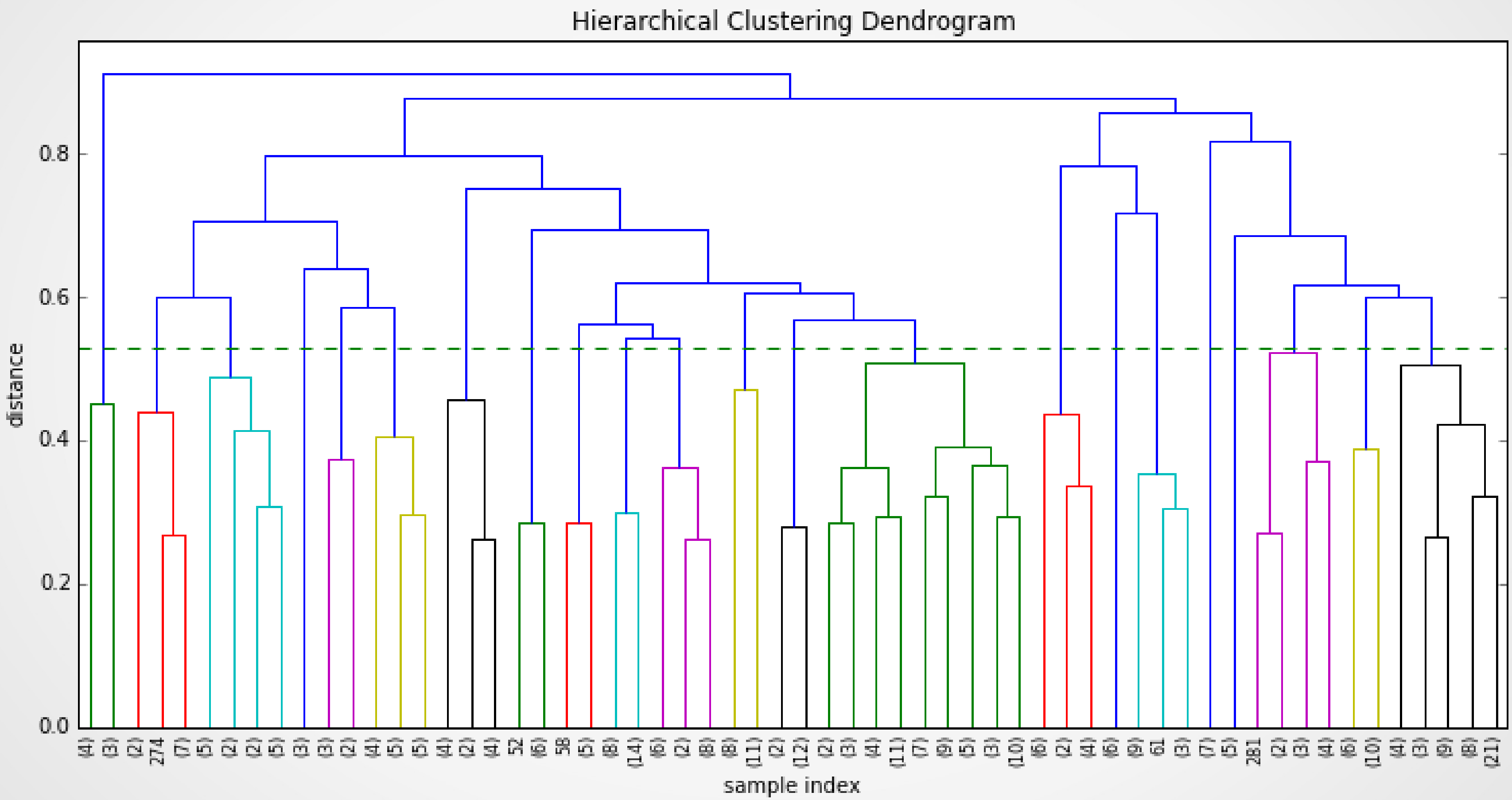
Methodology

- Steps
  - Web scraping (Selenium)
  - Corpus construction (NLTK, BeautifulSoup, ElementTree)
  - Text processing (Sklearn: tf/idf scaling, LSA)
  - Clustering (Scipy: agglomerative clust.)
  - Social Network Analysis and Visualization (Networkx, Cytoscape)

Methodology

- Output
  - A corpus of SOF posts including the text of the post, and the users who contributed to that post.
  - Top-voted posts on security.
  - Corpus size: 315 posts; 441253 tokens
  - A bipartite network of StackOverflow posts (as documents) and users
  - Posts are classified according to their latent semantic similarities into semantic classes (22)
- **Question**: Can we detect a community structure in this network with respect to the semantic class attribute?

Methodology



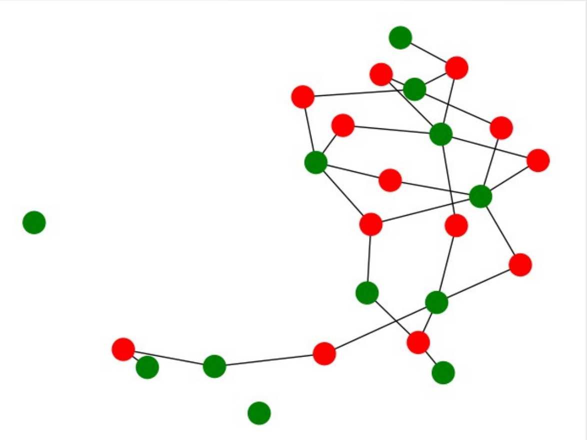
Methodology

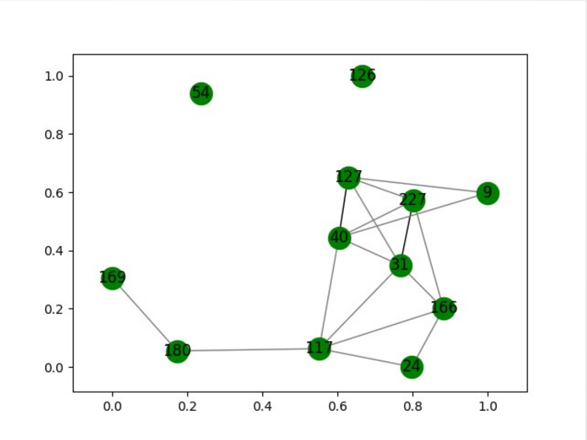
- Contents of the groups: One outlier, two primary clusters corresponding to client-side and server-side issues.

OUTLIER	1	SQL servers
SERVER-SIDE	2-3	asp-net, permissions, root
	4-6	Mobile, oath, authentication, securestring
	7	Java, Iframe, jquery
	8	Servers
	9-14	Users, passwords, encryption, databases, SQL injection
CLIENT-SIDE	15	Session security (cookies, remember user, etc.)
	16	web browsers: json
	17	web browsers: http, chrome firefox
	18	SSL, certificates
	19	Encryption
	20-22	a bit of everything (applications?), Java, vulnerabilities

Methodology

- Explore social links among posts
  - Use Information about posts and common users to build a bipartite network.
  - 721 users; 2219 edges
- Problem: community detection in (bipartite) networks (Fortunato 2010)
- Solution: Map the network onto one of its partitions
  - Classify nodes according to semantic class
  - Run clustering tests on it.
  - Example: Semantic class 15.





Network analysis

- Evaluate clustering
- Attribute Assortativity (Newman 2003)
  - Are nodes mostly connected to similar nodes?
  - Assortativity coefficient for semantic class: 0.0384
- Intra-cluster density: A community will have:

  a. more internal than external edges ($d_i - d_e > 0$)

  b. higher relative density than the whole network ($d_i - d > 0$)

Network analysis

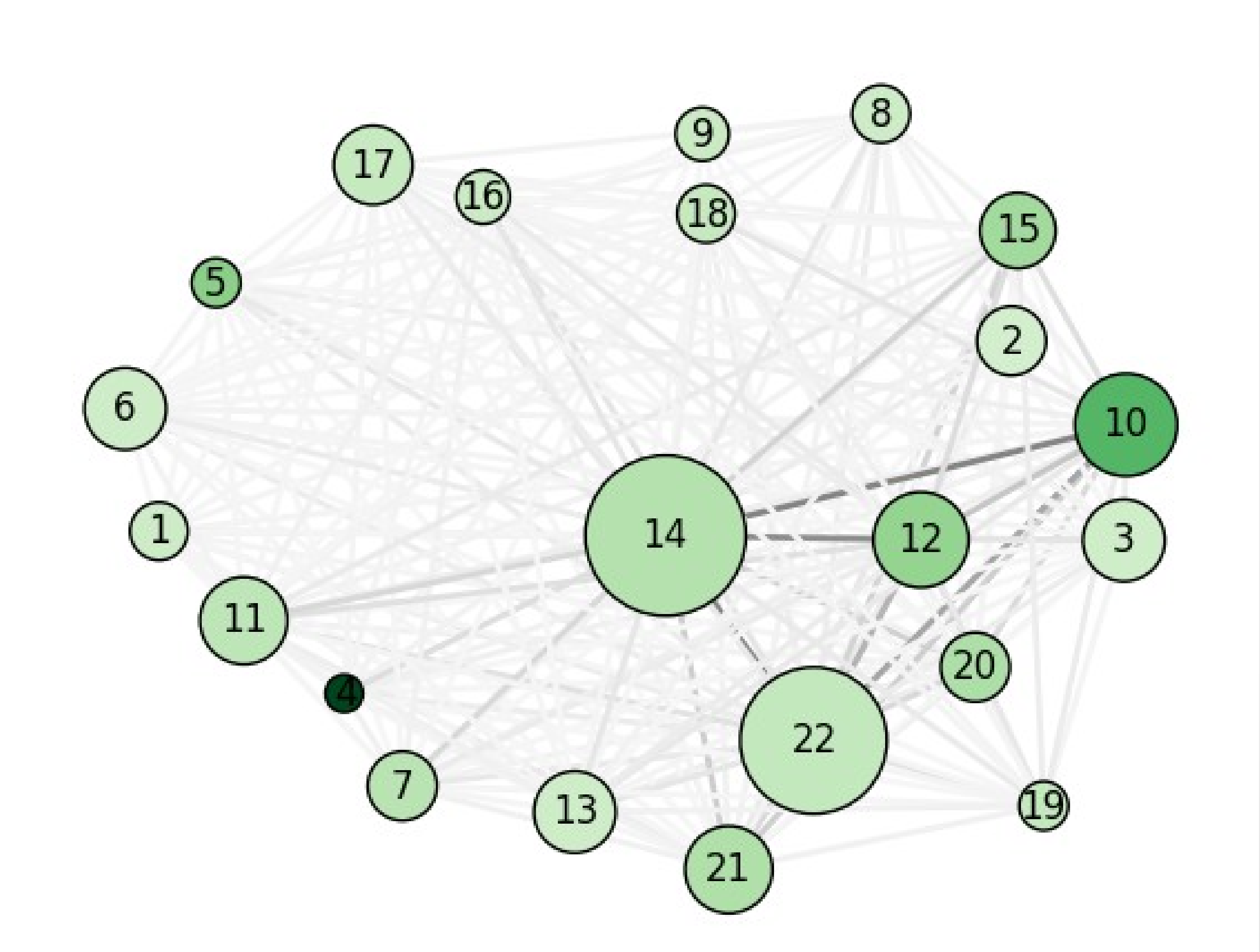
- Group density (communities indicated by * or **)

GROUP	#n	#e	d(i)	d(e)	d(i)-d(e)	d(i)-d
1	7	1	0.048	0.053	-0.005	-0.048
2	10	0	0.000	0.021	-0.021	-0.096
3	14	2	0.022	0.036	-0.014	-0.074
**4	3	4	1.333	0.154	1.179	1.237
**5	5	4	0.400	0.072	0.328	0.304
6	14	3	0.033	0.025	0.008	-0.063
*7	10	7	0.156	0.070	0.085	0.060
8	7	1	0.048	0.041	0.006	-0.048
9	6	1	0.067	0.007	0.060	-0.029
**10	22	145	0.628	0.134	0.493	0.532
*11	16	16	0.133	0.071	0.062	0.037

GROUP	#n	#e	d(i)	d(e)	d(i)-d(e)	d(i)-d
**12	19	59	0.345	0.124	0.221	0.249
13	14	4	0.044	0.052	-0.008	-0.052
*14	54	274	0.191	0.101	0.090	0.095
**15	12	19	0.288	0.107	0.181	0.192
16	6	1	0.067	0.078	-0.012	-0.029
17	13	7	0.090	0.035	0.054	-0.006
18	7	2	0.095	0.069	0.026	-0.001
19	5	1	0.100	0.119	-0.019	0.004
**20	10	10	0.222	0.110	0.113	0.126
**21	16	25	0.208	0.084	0.124	0.112
*22	45	100	0.101	0.091	0.010	0.005

Network analysis

- Graph
  - Size: # posts
  - Shade: density
  - Edge color: weight



Network analysis

- Evaluate clustering
  - Hypothesis 1 cannot be rejected, but it is not as strong as we would like.
  - There is a core component, connected by heavy weight edges: Not all users are the same?
  - There is evidence for clusters (communities) forming around posts with the same value for the semantic attribute, but more for some groups than others.

Linguistic and social factors

- Why do people take the time to contribute to PVCs like StackOverflow?

  - Immediate self-benefit: A high reputation (intangible remuneration) can be leveraged to land a good job or advance one's career

  - Triumph of the commons: Benefits are bestowed to the members of a group if the group itself thrives; members forgo of personal benefits to reap the collective benefits.

Linguistic and social factors

- Interpretation of results: dual composition of PVCs
  - A PVC may have a core of users that tend to the commons and reap collective benefits (“wardens”).
  - There may also be a swarm of users ("poachers") that ony contribute to the site occasionally for individual benefit.
  - Some content may be more attractive to one class of user than another.
  - Further research: Are both types of users necessary in the ecosystem of the thriving PVC?

Conclusions

- We have built a copus of SOF posts
- We classified the posts according to content and we mapped them onto a network according to social links
- We found that posts form communities based on their content, but in a non-uniform way.
- The findings support a view of PVCs in which members may have dual motives to participate.
- Using social links to enhance the extration of semantic information 9Yazdani & Popescu-Belis 2013).

Selected References

Fortunato, Santo (2010). Community detection in graphs. *Physics Reports* 486 75–174

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.

Newman, M. E. J. (2003). Mixing patterns in networks. *Physical Review* E 67, 026126 1-13

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.

Yazdani, Majid & Andrei Popescu-Belis (2013). Computing text semantic relatedness using the contents and links of a hypertext encyclopedia. *Artificial Intelligence* 194: 176–202