Risk detection in OSS teams with NLP

R. Aranovich, P.T. Devanbu, V. Filkov – UC Davis
<raranovich@ucdavis.edu>, <filkov@cs.ucdavis.edu>

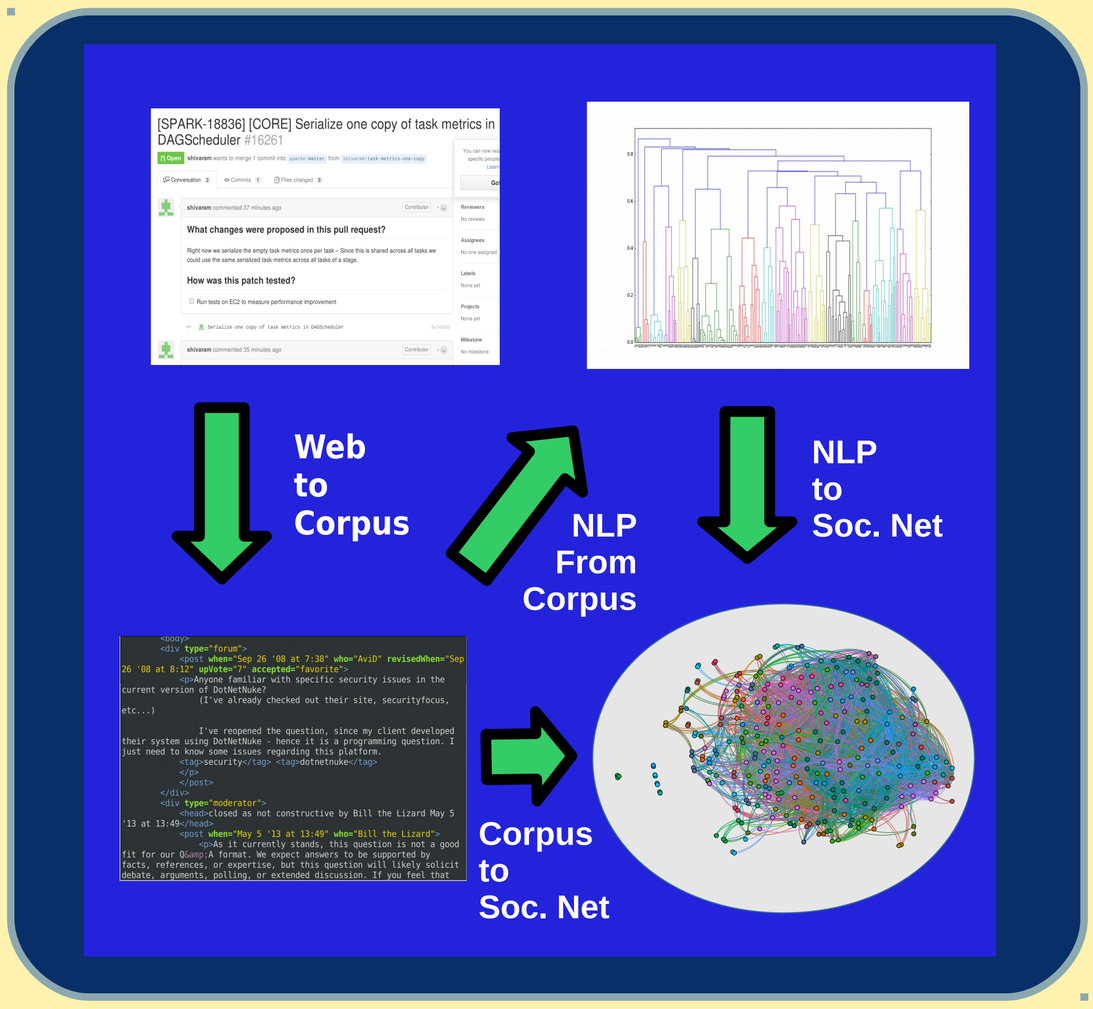
Project Constitution, Goals, Focus, and Scope

Collaboration between computer scientists (software engineering) and social scientists (linguistics) to develop a way to identify potential system vulnerabilities introduced in the process of developing Open Source Software (OSS)

Goal: make code more secure by flagging developers more likely to introduce vulnerabilities in a program.

Hypothesis: "high risk" developers display patterns of linguistic behavior (in their use of both natural language and code) that are different from those of seasoned (i.e. "low risk") developers.

Outcome: Detection of those linguistic patterns are used to evaluate the comparative risk level of a developer.



Approach

- Discover stylistic clusters in virtual communication networks Using Natural Lang. Processing and Machine learning
- Apply social network theory to predict clusters based on group relations
- Build model on the assumption that divergent linguistic norms in a community are the outside manifestation of individual degree of embedding.
- Correlate linguistically determined embedding with risk factor of code

Step 1: Finding a linguistic baseline

Software developers in on-line communities, especially those who focus on cybersecurity, have specific communicative patterns. To find a baseline, we assembled a corpus of posts from StackOverflow, hand-tagged and annotated it to preserving the structure of on-line communicative interactions.

Step 2: Develop tools to process mixed English/Code textual sources

Applying NLP techniques to the kind of mixed language used by developers poses a technical challenge: how to process code snippets (incomplete bits of a program found inside comments and Q/A posts). We are developing a parser with that capability.

English and Code are Not So Different

Developers mix code and English in their communications in a way that supports the "naturalness of code" thesis (programming languages have many features in common with natural languages)

Social Network Theory and CMC

Development of a schema for annotating corpora containing mixed languages (code/English) and also for keeping track of the internal social structure of the community engaged in communication over the web, using XML tags.

Interested in meeting the PIs? Attach post-it note below!

National Science Foundation
WHERE DISCOVERIES BEGIN

The 3rd NSF Secure and Trustworthy Cyberspace Principal Investigator Meeting
January 9-11, 2017
Arlington, Virginia