A Case Study of Open Source Software

Measuring the Effect of Social Communications on Individual Working Rhythms

Qi Xuan, Mohammad Gharehyazie, Premkumar T Devanbu, Vladimir Filkov
Department of Computer Science, University of California, Davis
Email: qxuan@ucdavis.edu
Rhythm in Music

Fryderyk Fanciszek Chopin
1810-1849

Maksim Mrvica
1975-

Slow

Fast
Rhythm in OSS

Files

Rhythms

Productivities, Responsibilities, etc

Use/Evaluate

Commit

Bookkeeper

Committer #1

Committer #2
Work Vs. Talk: Negative effects

Few words, many deeds.
Work VS Talk: Positive effect

Evaluation-Response Mechanism

Commit files

Review files

Report bugs

Fix bugs

With great power comes great responsibility.
Which is true?

Do the communication activities impede committing activities or do they accelerate them?
Our contribution

Propose two methods:
Method #1: Network based method
Method #2: Time-series based method

to macroscopically and microscopically quantify the relationship between committing rhythm and communication activities of developers based on the real OSS data collected from Apache Software Foundation.
# Data description

<table>
<thead>
<tr>
<th>Project</th>
<th>$N_T$</th>
<th>$N_D$</th>
<th>$N_F$</th>
<th>$\langle k_{out} \rangle$</th>
<th>$\langle k_{in} \rangle$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accumulo</td>
<td>75</td>
<td>5</td>
<td>1622</td>
<td>3.5</td>
<td>3.5</td>
</tr>
<tr>
<td>Mahout</td>
<td>552</td>
<td>16</td>
<td>5123</td>
<td>4.4</td>
<td>4.4</td>
</tr>
<tr>
<td>Lucene</td>
<td>2148</td>
<td>41</td>
<td>6674</td>
<td>4.2</td>
<td>4.2</td>
</tr>
<tr>
<td>Nutch</td>
<td>862</td>
<td>16</td>
<td>3072</td>
<td>3.4</td>
<td>3.4</td>
</tr>
<tr>
<td>Derby</td>
<td>1128</td>
<td>35</td>
<td>6563</td>
<td>5.6</td>
<td>5.6</td>
</tr>
<tr>
<td>Ode</td>
<td>377</td>
<td>18</td>
<td>11006</td>
<td>3.8</td>
<td>3.8</td>
</tr>
<tr>
<td>Openejb</td>
<td>179</td>
<td>38</td>
<td>43960</td>
<td>5.8</td>
<td>5.8</td>
</tr>
<tr>
<td>Log4php</td>
<td>88</td>
<td>9</td>
<td>1409</td>
<td>2.1</td>
<td>2.1</td>
</tr>
<tr>
<td>Wicket</td>
<td>540</td>
<td>24</td>
<td>48045</td>
<td>5.5</td>
<td>5.5</td>
</tr>
<tr>
<td>Log4j</td>
<td>540</td>
<td>19</td>
<td>5519</td>
<td>2.3</td>
<td>2.3</td>
</tr>
<tr>
<td>Bookkeeper</td>
<td>32</td>
<td>3</td>
<td>407</td>
<td>3.0</td>
<td>3.0</td>
</tr>
</tbody>
</table>

702 developers in 31 projects
From Apache Software Foundation on March 24th, 2012
Definition of Rhythm

Time-series of committing activities:

List of inter-activity time intervals:

\[ \Delta t_i = t_{i+1} - t_i, \quad i = 1, 2, \ldots, M - 1 \]

Average inter-activity time interval:

\[ \langle \Delta t \rangle = \frac{\sum_{i=1}^{M-1} \Delta t_i}{M - 1} = \frac{t_M - t_1}{M - 1} \]
RQ #1: Whether the developers with higher social status have faster or slower committing rhythms than those with lower social status?
Two Cases

Case I: The developers are divided into two groups according to their *incoming degrees*.

\[ k_{in} \leq 50 \quad \text{or} \quad k_{in} > 50 \]

Case II: The developers are divided into two groups according to their *outgoing degrees*.

\[ k_{out} \leq 50 \quad \text{or} \quad k_{out} > 50 \]
## Results

<table>
<thead>
<tr>
<th>T-test</th>
<th>$k \leq 50$</th>
<th>$k &gt; 50$</th>
<th>T-value</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case I</td>
<td>103.0 (h)</td>
<td>45.5 (h)</td>
<td>20.4</td>
<td>$p &lt; 10^{-6}$</td>
</tr>
<tr>
<td>Case II</td>
<td>111.2 (h)</td>
<td>48.5 (h)</td>
<td>22.3</td>
<td>$p &lt; 10^{-6}$</td>
</tr>
</tbody>
</table>

Higher social status indicate faster committing rhythms.
Time-Series Based Method

RQ #2: Is there an evaluation-response mechanism in OSS projects so that committing and communication activities accelerate each other?
Evaluation & Response Latency

Submit codes  Evaluations  Response

Code reviews and problem reports

Evaluation latency: \[ \tau^E = t_2 - t_1 \]

Response latency: \[ \tau^R = t_{k+1} - t_k \]
Random Time-Series

(a) Real time-series

(b) Randomly reorder time intervals

(c) Random time-series
Results

<table>
<thead>
<tr>
<th></th>
<th>T-test</th>
<th>simulated</th>
<th>Real</th>
<th>T-value</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Evaluation</td>
<td>89.9 (h)</td>
<td>58.5 (h)</td>
<td>8.21</td>
<td>$p &lt; 10^{-6}$</td>
<td></td>
</tr>
<tr>
<td>Response</td>
<td>97.0 (h)</td>
<td>57.2 (h)</td>
<td>9.92</td>
<td>$p &lt; 10^{-6}$</td>
<td></td>
</tr>
</tbody>
</table>

Communication and committing activities accelerate each other.
Limitations & Future Works

- **Limitations**: In this work, we just use the time-series of committing and communication activities to estimate the evaluation and response latencies, which needs to be further checked by textural analysis of emails and codes.

- **Future works**: the network and time-series methods can be used together in order to reveal whether the individuals response at different rhythms for the evaluations from others of different social status.
Thanks!