When your social position predicts your success
Lessons from Open Source communities and citations

Ingo Scholtes
Chair of Systems Design
ETH Zürich

1 in collaboration with René Pfizner, Emre Sarigöl, Marcelo Serrano Zanetti, Antonios Garas and Frank Schweitzer
Socio-technical systems

data-driven modeling
When your social position predicts your success

Ingo Schultes

Chair of Systems Design
www.sg.ethz.ch

Software engineering
Software defects: bug tracking

information overload!
<table>
<thead>
<tr>
<th></th>
<th>Firefox</th>
<th>Thunderbird</th>
<th>Eclipse</th>
<th>NetBeans</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Start date</strong></td>
<td>April 2002</td>
<td>January 2000</td>
<td>October 2001</td>
<td>January 1999</td>
<td></td>
</tr>
<tr>
<td><strong>Bug reports</strong></td>
<td>112,968</td>
<td>35,388</td>
<td>356,415</td>
<td>210,921</td>
<td>715,692</td>
</tr>
<tr>
<td><strong>Changes</strong></td>
<td>1,068,070</td>
<td>313,957</td>
<td>2,594,385</td>
<td>1,875,878</td>
<td>5,852,290</td>
</tr>
<tr>
<td><strong>Valid</strong></td>
<td>21 %</td>
<td>24 %</td>
<td>74 %</td>
<td>63 %</td>
<td>57 %</td>
</tr>
<tr>
<td><strong>Faulty</strong></td>
<td>79 %</td>
<td>76 %</td>
<td>26 %</td>
<td>37 %</td>
<td>43 %</td>
</tr>
</tbody>
</table>
When your social position predicts your success

role of social structures in the bug handling process?
From bug interactions …
When your social position predicts your success

to evolving collaboration networks

we observe how team structures influence team performance

Hypothesis: position of bug reporter in the monthly collaboration network preceding the time of report is indicative for outcome of the bug handling process.
**Hypothesis:** position of bug reporter in the monthly collaboration network preceding the time of report is indicative for outcome of the bug handling process

**YES, it is!**
Prediction method

1. in largest connected component?
2. eigenvector centrality
3. closeness centrality
4. betweenness centrality
5. k-coreness
6. total degree
7. in-degree
8. out-degree
9. local clustering coefficient

Support Vector Machine

valid or faulty?

training set
## Prediction results

<table>
<thead>
<tr>
<th></th>
<th>Firefox</th>
<th>Thunderbird</th>
<th>Eclipse</th>
<th>NetBeans</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Valid</strong></td>
<td>21 %</td>
<td>23 %</td>
<td>74 %</td>
<td>63 %</td>
</tr>
<tr>
<td><strong>Precision</strong></td>
<td>82.5 %</td>
<td>90.3 %</td>
<td>88.7 %</td>
<td>78.9 %</td>
</tr>
<tr>
<td><strong>Recall</strong></td>
<td>44.5 %</td>
<td>38.9 %</td>
<td>91.0 %</td>
<td>87.0 %</td>
</tr>
<tr>
<td><strong>F-Score</strong></td>
<td>0.58</td>
<td>0.54</td>
<td>0.89</td>
<td>0.83</td>
</tr>
</tbody>
</table>

- better prediction than previous approaches on same data
- the social network matters!
When your social position predicts your success

Collaboration–Citation–Multiplex

Chair of Systems Design
www.sz.ethz.ch

Lorentz Center Workshop 2014, Leiden
09.04.2014
Collaboration–Citation-Multiplex

Predicting Scientific Success Based on Coauthorship Networks

Emre Sarıgöl, René Pfitzner*, Ingo Scholtes, Antonio Garas, Frank Schweitzer
Chair of Systems Design, ETH Zurich
CH-8092 Zürich, Switzerland
*rpfitzner@ethz.ch

Abstract

We address the question to what extent the success of scientific articles is due to social influence. Analyzing a data set of over 300,000 publications from the field of Computer Science, we study how centrality in the coauthorship network differs between authors who have highly cited papers and those who do not. We further show that a machine learning classifier, based only on coauthorship network centrality measures at time of publication, is able to predict with high precision whether an article will be highly cited five years after publication. By this we provide quantitative insight into the social dimension of scientific publishing – challenging the perception of citations as an objective, socially unbiased measure of scientific success.

1 Introduction

Quantitative measures are increasingly used to evaluate the performance of research institutions, departments, and individual scientists. Measures like the absolute or relative number of published research articles are frequently applied to quantify the productivity of scientists. To measure the impact of research, citation-based measures like the total number of citations, the number of citations per published article or the h-index [5], have been proposed. Proponents of such citation-based measures or rankings argue that they allow to quantitatively and objectively assess the quality of research, thus encouraging their use as simple proxies for the success of scientists, institutions or even whole research fields. The intriguing idea that by means of citation metrics the task of assessing research quality can be “outsourced” to the collective intelligence of the scientific community, has resulted in citation-based measures becoming increasingly popular among research administrations and governmental decision-makers. As a result, such measures
Evolving coauthorship networks

\[ \alpha \]
Ralph Alpher

\[ \beta \]
Hans Bethe

\[ \gamma \]
George Gamow

Image credit: Wikimedia Commons
Evolving coauthorship networks

Hypothesis: position of authors in the coauthorship network up to two years before time of publication is indicative for citation success of a paper

| Hypothesis: | position of authors in the coauthorship network up to two years before time of publication is indicative for citation success of a paper |


YES, it is!
Top 10% paper?

- Network metrics capturing social position of authors
  - (maximum eigenvector centrality, betweenness centrality, degree, k-core, ...)

Random Forest Classifier

data set
- 108,758 CS publications
- 160,891 researchers
- 5.3 Million collaborations

Time

1946

1948

When your social position predicts your success

Go Scholtes

Top 10% paper?
Social position correlates with future success

Results (CS)

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>60 %</td>
</tr>
<tr>
<td>Recall</td>
<td>18 %</td>
</tr>
<tr>
<td>F-Score</td>
<td>0.28</td>
</tr>
</tbody>
</table>

- six times better than random guess
- social position of authors alone sufficient to predict citation success of 2,000 out of 10,800 top papers
- data mining + network analysis + predictive analytics
- highlights influence of social structures in information filtering
- reputation is multi-dimensional and time-varying concept

Social position correlates with future success

Results (CS)

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>60 %</td>
</tr>
<tr>
<td>Recall</td>
<td>18 %</td>
</tr>
<tr>
<td>F-Score</td>
<td>0.28</td>
</tr>
</tbody>
</table>

- six times better than random guess
- social position of authors alone sufficient to predict citation success of 2,000 out of 10,800 top papers
- data mining + network analysis + predictive analytics
- highlights influence of social structures in information filtering
- reputation is multi-dimensional and time-varying concept
Closing the loop?

- feedback between collaboration and citation dynamics?
- implications for ranking of scientists?
When your social position predicts your success


http://www.sg.ethz.ch

+Chair of Systems Design

@ingo_s

ischoltes@ethz.ch