Oracle, Where Shall I Submit My Papers?

June 2018 @ UCAS

Dongwon Lee

Penn State / IST
The progress of science as a mission for discovery

Producers of discovery

Distributors of discovery

Consumers of discovery

RESEARCHERS — investigation and communication

Pre-publication peer-review filters — fraud detection and facilitation of effective communication (e.g. improving manuscript quality)

PUBLISHERS — storage and dissemination

Post-publication peer-review tags — merit ranking (e.g. citation) metrics for both the product (e.g. article) and the producer (author reputation)

SOCIETY — learning and application

Source: https://science-open-reviewed.com/?page_id=8
The Evolution of Intellectual Freedom

Before grad school
Grad student
Assistant professor
Tenured professor
Emeritus professor

I'm going to research whatever I want!
I'm going to research whatever my professor wants!
I'm going to research whatever my tenure committee wants!
I'm going to research whatever my grant committee wants!
I'm going to research whatever I want!

“Research in Peace”

WWW.PHDCOMICS.COM

Jorge Cham © 2011
Differences in Disciplines

- **Pure Sciences (eg, Math, Physics)**
  - Pre-print at Arxiv.org
  - Rigorous reviews for journals
  - Huge flagship conference (ICM attracted ~4000)

- **Social Sciences**
  - Often value journals > conferences
  - Conferences are mostly for gathering or short abstract based screening
  - Rigorous reviews for journals

- **Computational Sciences (CSE, iSchool)**
  - Peer-reviewed conferences
  - Top conferences have 5-15% acceptance rate
  - Specialized and small conferences (attendance of 500+)
  - Often value conferences > journals
How three MIT students fooled the world of scientific journals
A decade later, CSAIL alumni reflect on their paper generator and reveal a new fake-conference project.

Watch Video

Adam Conner-Simons | CSAIL
April 14, 2015
Rooter: A Methodology for the Trust of Access Points and Routers

Jeremy Stribling, Daniel Aguayo and Max "Novice" Brown

Abstract

Many physicists would agree that, had it not been for the invention of the web browsers, the evaluation of web servers might never have occurred. In fact, few hackers worldwide would disagree that the essential unification of voice-over-IP and public-key cryptography has led to the world as we know it. In order to solve this riddle, we confirm that if a browser can be made stochastic, cacheable, and interposable.

1. Introduction

The rest of this paper will motivate the work in context. We address this obstacle by tauting autonomous to-analog conversion oriented languages designed. Along the way, we concentrate on...
Motivation

- Digital Libraries provide effective recommendation and filtering tools
- Computer Science is unique in its publication practice:
  - often value conferences > journals
  - constantly increasing number of conferences
- Existing bibliometrics are becoming inadequate to measure the quality of conferences in CS
Why inadequate

- Inapplicable for emergent conferences: historical citation statistics are not readily available

- *Delay* for well-established conferences: citation statistics takes time to accumulate
  - major database conferences and journals between 1994 and 2003 had most citations reach back five and more years [Rham & Thor, 2005]

- Require disciplinary knowledge: unless it is a well-known venue in a researcher’s familiar domain, he/she may not be aware of its citation statistics
How to assess venue quality?

- What venues are reputable?
- What can be said about questionable ones?

Scientific conference falls for gibberish prank

A bunch of computer-generated gibberish masquerading as an academic paper has been accepted at a scientific conference in a victory for pranksters at the Massachusetts Institute of Technology (MIT).

Jeremy Stirling said that he and two fellow MIT graduate students questioned the standards of some academic conferences, so they wrote a computer program to generate research papers complete with nonsensical text, charts and diagrams.

The trio submitted two of the randomly assembled papers to the World Multi-conference on Systemics, Cybernetics and Informatics (WMSCI), scheduled to be held July 10-13 in Orlando, Florida.

To their surprise, one of the papers - "Rooter: A Methodology for the Typical Unification of Access Points and Redundancy" - was accepted for presentation.

ABC Online

Scientific conference falls for gibberish prank. 15/04/2005. ABC News Online

[This is the print version of story http://www.abc.net.au/news/newsitem.cfm?20010404145733.html]

Last Update: Friday, April 15, 2005, 8:19am (ABST)

http://pdos.csail.mit.edu/scigen/
Question

Can we automatically identify the **reputable** (and low quality) conferences among the hundreds of Call For Papers (CFP) each year?

- Desiderata for our system
  - Large-number of venues per year ⇔ must be scalable
  - Automatic detection ⇔ little manual processing
  - False positives >> false negatives

Number of CFPs posted to DBWorld per year
CFP Example

Program Chairs

David Evans, University of Virginia
Tal Malkin, Columbia University
Dongyan Xu, Purdue University

Program Committee

Sadia Afroz, UC Berkeley / ICSI
Gail-Joon Ahn, Arizona State University
Ehab Al-Shaer, University of North Carolina Charlotte
Elias Athanasopoulos, University of Cyprus
Foteini Baldimtsi, George Mason University
David Basin, ETH Zurich
Adam Bates, University of Illinois at Urbana-Champaign
Lujo Bauer, Carnegie Mellon University
Konstantin Beznosov, University of British Columbia
Karthikeyan Bhargavan, INRIA
Alex Biryukov, University of Luxembourg
Jeremiah Blocki, Purdue University
Elette Boyle, IDC Herzliya
Levente Buttyán, CrySyS Lab, BME
Juan Caballero, IMDEA Software Institute

Marcel Keller, University of Bristol
Aggelos Kiayias, University of Edinburgh
Taesoo Kim, Georgia Tech
Yongdae Kim, KAIST
Engin Kirda, Northeastern University
David Kotz, Dartmouth
Farinaz Koushanfar, UC San Diego
Ralf Küsters, University of Stuttgart
Andrea Lanzi, University of Milan
Byoungyoung Lee, Purdue University
Wenke Lee, Georgia Tech
Brian N. Levine, University of Massachusetts Amherst
Zhichun Li, NEC Labs
Zhou Li, RSA
David Lie, University of Toronto
Yao Liu, University of South Florida
Matteo Maffei, TU Vienna
Mohammad Mahmody, University of Virginia
Z. Morley Mao, University of Michigan
Ivan Martinovic, University of Oxford
Michelle L. Mazurek, University of Maryland
Jonathan McCune, Google
Jonathan M. Smith, University of Virginia / Navy
Candidate features for conference rating

- Good vs. bad venues
  - Citation counting (eg, Impact Factor)
  - Acceptance rate
  - Reputation (eg, ACM, IEEE, etc.)
  - Longevity

- At the end, none satisfies all of our desiderata. Need something else…
Qualities of venues are closely correlated with quality of the Program Committee (PC) members

- PC member list can be readily obtained from CFP ⇔ data extraction + data cleaning
- Each CFP has only finite number of PCs ⇔ scalability
- Examine quality of PC members w.r.t heuristics:
  - Citation counting, productivity, social network attributes, …
Methodology

Given a venue with CFP

- Extract the PC list
- Map the authors in the PC list to our social network of CS community
- Evaluate the PC list w.r.t. different features
- Use a trained classifier to classify the PC list into one of the classes Reputable/Not
Data collection

- CFPs posted to the DBWorld mail list:
  - 2,979 CFPs (feb – may 2006)
  - 16,147 distinct PC members

- Publication metadata in ACM DL
  - 609K authors, 770K articles published in 1950-2004
  - 1.2M edges on the co-authorship network

- Ground truth about conference quality: CS-Conference-Ranking.org (CSCR)
  - Reputable conferences: 576 CFPs labeled as R
  - Disreputable (common) conferences: 2,403 CFPs labeled as C
Features for classification

- Number of PC members
- Number of publications of PC members
- Number of co-authors of PC members
- Closeness centrality of PC members
- Betweenness centrality of PC members

Source: https://toreopsahl.com/tnet/weighted-networks/node-centrality/
Average # of PCs
Average # of papers by PCs

![Graph showing the average number of papers by PCs]
Average # of co-authors

![Chart showing the average number of co-authors and the probability of R. The chart includes a bar graph and a line graph, indicating the distribution of the average number of co-authors across different ranges.]
Average betweenness

The diagram illustrates the distribution of conferences with different levels of average betweenness. The x-axis represents the average betweenness values, ranging from 4.00E-05 to 0.0004, and the y-axis shows the fraction of conferences ranging from 0% to 60%. The graph also displays the probability of R on the right y-axis, ranging from 0 to 0.7.
Classification with Decision Tree

PC has feature A?
PC has feature B?
PC has feature C?
PC has feature D?

Reputable venue
Disreputable venue

training set

Reputable?
Performance for combining all features

- **Decision Tree (C4.5)**

<table>
<thead>
<tr>
<th>class</th>
<th>precision</th>
<th>recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>0.877</td>
<td>0.965</td>
</tr>
<tr>
<td>R</td>
<td>0.751</td>
<td>0.434</td>
</tr>
</tbody>
</table>

- **Bagging**

<table>
<thead>
<tr>
<th>class</th>
<th>precision</th>
<th>recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>0.899</td>
<td>0.979</td>
</tr>
<tr>
<td>R</td>
<td>0.859</td>
<td>0.540</td>
</tr>
</tbody>
</table>

- **Boosting**

<table>
<thead>
<tr>
<th>class</th>
<th>precision</th>
<th>recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>0.938</td>
<td>0.964</td>
</tr>
<tr>
<td>R</td>
<td>0.831</td>
<td>0.733</td>
</tr>
</tbody>
</table>
Detect low-quality conferences

- Data sources:
  - Hand selected 18 low-quality conferences from fakeconferences.org: 18 CFPs, labeled as LQC
  - Common conferences: 2,403 CFPs, labeled as C

- Classified using the combination of
  - # of PC
  - # of publication of PC
  - # of co-authors of PC
  - Closeness centrality of PC
  - Betweenness centrality of PC
Avg. closeness of PCs (C & LQC)
C vs. LQC: Classification results

- Naïve C4.5 classifier
  - LQC Precision: 0.996
  - LQC Recall: 0.998
- Correctly judged 99.38% instances
- False positive rate is 0.002
- Detected two low-quality conferences that are not part of the original LQC list
PSU’s MIT Emulation

- Apr. 10, 2006, we generated 3 bogus papers using MIT SCIgen software:
  - P1 by Ethan Patel
  - P2 by Simon R. Hathaway
  - P3 by Richard Zhang
Measuring paper authenticity

- Indiana’s Inauthentic Paper Detector says:
  - P1: 28.9% => inauthentic
  - P2: 61.5% => authentic
  - P3: 38% => inauthentic
LQC Evaluation

Conference A and B

- April 24 – May 1, 2006
  - P1 submitted to Conf A on April 24
  - P2 submitted to Conf B on April 26
  - P3 submitted to Conf A on May 1

- May 15, 2006
  - P1 and P2 accepted w/o reviews
  - P3 rejected w/o reviews
  - Asked for reviews or any rationale ⇔ no response so far
“Ethan Patel” made it!

Pradip Peter Dey, Ph.D.
Professor, SoMET
Department of Computer Science and Information Systems
School of Engineering and Technology
National University
11255 North Torrey Pines Road
La Jolla, CA 92037-1011
U.S.A.
Phone: 858-642-8486
Fax: 858-642-3489
E-Mail: pdey@nu.edu

4th International Conference on Computer Science and its Applications (ICCSA-2006)
June 27-29, 2006  San Diego, California, USA

Date: May 15, 2006
Paper#: ICCSA_The_1142
Title: The Impact of Client-Server Epistemologies on Software Engineering
Author: Ethan Patel
BST Research Institute, USA
ethan_patel@yahoo.com

Dear author,

We are pleased to inform you that your above mentioned paper has been accepted for presentation at the 4th International Conference on Computer Science and its Applications (ICCSA-2006) and for publication in the conference proceedings. Please submit the final camera-ready version of the paper electronically to pdey@nu.edu or through our website. In addition, please submit the completed registration form along with the copyright transfer form as soon as possible. The forms are available at the conference web site:

http://www.conferencehome.com

Please remember that at least one author must register for the conference.

Sent: Tue 2006-05-16 4:58 PM