Data Science for Community Managers

J. Manrique López de la Fuente
@jsmanrique
jsmanrique at bitergia dot com
https://speakerdeck.com/bitergia
Outline

Introduction

Open Development Communities

Open Development Analytics

Playing with GrimoireLab

Extras
Introduction

A bit about me
Why am I here?
Disclaimer
Hello, my name is Manrique and I am a community junkie

Involved in: HPCC, AsturLiNUX, HispaLiNUX, GPE, Maemo, Meego, Gnome, GDG, Mozilla, ...

Business & marketing developer in Bitergia, the software development analytics company
Manrique Lopez
@jsmanrique
Mobile & Open Source technologies, Friends, Surf and Live passionate. Bilbaino born in Gijón living in Fuenla and working in @bitergia
I am not a computer scientist, how many of you are?
I am not a data scientist, how many of you are?
Presentation focus in *open source* & *inner source* related communities
Open Development Communities

What is a community?
Community Management
“A community is commonly considered a social unit (a group of people) who share something in common, such as norms, values, identity, and often a sense of place that is situated in a given geographical area (e.g. a village, town, or neighborhood). Durable relations that extend beyond immediate genealogical ties also define a sense of community. People tend to define those social ties as important to their identity, practice, and roles in social institutions like family, home, work, government, society, or humanity, at large. Although communities are usually small relative to personal social ties (micro-level), "community" may also refer to large group affiliations (or macro-level), such as national communities, international communities, and virtual communities.

The word "community" derives from the Old French comuneté which comes from the Latin communitas (from Latin communis, things held in common)"
It’s about people
It’s about people who share
It’s about people who share something in common
Self-awareness - Governance - Transparency
Anatomy of a community

Alex Hillman
Governance

“Establishment of policies, and continuous monitoring of their proper implementation, by the members of the governing body of an organization. It includes the mechanisms required to balance the powers of the members (with the associated accountability), and their primary duty of enhancing the prosperity and viability of the organization”.

businessdictionary.com
/communities

Transparency to the community

Fairness

Transparency to third parties

Trust
The Community Manager

Do communities need to be managed?

Do communities need leadership?

Why is becoming so important?

How many of you are community managers, developers relationship (AKA DevRel), etc.?
Community Manager Responsibilities

Community health

Community productivity

Community visibility
“To measure is to know”

“If you can not measure it, you can not improve it”

Lord Kelvin

“Without data, you are just another person with an opinion”

W. Edwards Deming
“Human beings adjust behavior based on the metrics they're held against. Anything you measure will impel a person to optimize his score on that metric. What you measure is what you’ll get. Period”.

You Are What You Measure by Dan Ariely
Open Development Analytics

Transparency & objectivity matters

Data Sources
Tools
GrimoireLab
The community manager nightmare!

I need information
The community manager nightmare:

- Time is limited
- How to take decisions without reliable data?
Where is everything happening?
Data silos
Several interesting approaches

OpenHub

StackOverflow metrics

Stackalytics

GitTorrent, GitHub Archive

GitHub Archive + Google BigQuery

github3.py
Architecture

Data sources

Config. file

Arthur

Sorting Hat DB

Enrichment

ES raw data

ES rich data

Perceval

GrimoireLab

Kibana based dashboards

ES API

Reports

Bitergia

/grimoirelab
PERCEVAL
Sir Perceval goes on the quest to retrieve and gather data from git, GitHub, Bugzilla, JIRA, Gerrit, inbox, pipermail, StackExchange, Discourse, etc.
- Check Perceval repository

SORTING HAT
Tool to manage people identities information, to merge multiple person identities across different data sources, enrich profile information with affiliation information, etc.
- Check Sorting Hat repository

GRIMOIREELK
Playground for testing the whole set of tools as a platform for software development analytics. It's is a prototype of the Grimoire Open Development Analytics platform.
- Check GrimoireELK repository

KIBITER
Custom fork of Kibiter to work on new ideas for metrics & data visualization to be used by GrimoireLab Panels.
- Check KIBiter repository

PANELS
Set of pre-defined widgets and dashboard templates to visualize Elasticsearch indexes generated by GrimoireELK with Kibiter.
- Check Panels repository

ARThUR
King Arthur commands his loyal knight Sir Perceval managing the tasks to retrieve data for analysis. It manages data incremental update, parallel downloading, etc.
- Check Arthur repository

Bitergia
grimoirelab.github.io
Some features

Drill down
Time frame selection
Sharing / embedding
Data export (CSV...)
Query API (ElasticSearch)

Users can create custom widgets and panels
Easy validation
Links to real artifacts (commits, tickets, etc.)
Search box
Playing with GrimoireLab
Let’s start..

$ git clone http://.../perceval.git
$ sudo python3 setup.py install
...
$ git clone http://.../grimoireelk.git
$ elasticsearch/bin/elastic &
$ kibana/bin/kibana &
GrimoireELK/utils$ python3 ./p2o.py --enrich --index git_yarn -e http://localhost:9200
--no_inc --debug git https://github.com/yarnpkg/yarn.git
...
2016-10-14 14:09:32,392 Total items enriched 897
2016-10-14 14:09:32,392 Done git
2016-10-14 14:09:32,392 Enrich backend completed
2016-10-14 14:09:32,393 Finished in 0.75 min
GrimoireELK/utils$ python3 ./p2o.py --enrich --index github_yarn -e http://localhost:9200 --no_inc --debug github -t *** --owner yarnpkg --repository yarn
...
2016-10-14 14:20:19,736 Total items enriched 900
2016-10-14 14:20:19,736 Done github
2016-10-14 14:20:19,736 Enrich backend completed
2016-10-14 14:20:19,738 Finished in 7.20 min
$
GrimoireLab is free, open source software for software development analytics. It allows you to retrieve data from many kinds of systems with information related to software development, and produce analysis and visualizations with it.

GrimoireLab supports more than 20 different kinds of data sources, from git repositories or GitHub projects, to issue trackers such as Jira or Bugzilla, including messaging systems such as IRC, Slack or mailing lists, or other types of systems such as StackOverflow or Jenkins.

With the tools in the GrimoireLab toolset you can retrieve data, store it in databases, analyze it, and produce nice dashboards for interactively visualizing the resulting information.
Extras

Some analysis and metrics provided with Grimoire Lab

Network analysis
Dependency
Geographical analysis
Demography
Gender diversity
Contributors review
Dealing with issues
Contributors funnel
Open Source Kibana network visualization plug-in:
https://github.com/dlumbrer/kbn_network
Onion model
- 7 core
- ~40 regular
- ~85 casual

ASF Pony factor
- Pony factor: 1

Bitergia Elephant factor
- Elephant factor: 2

Bitergia Zapata factor
- Linux Kernel Zapata factor ~ 200

Bitergia United Fruit Company factor
- Linux Kernel UFCo factor ~ 10

Linux kernel ownership analysis: [linux.biterg.io](http://linux.biterg.io)
### Open Containers Initiative Demography panel

#### List of Authors

<table>
<thead>
<tr>
<th>Author</th>
<th>Count</th>
<th>Last commit</th>
<th>Last commit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sebastian</td>
<td>103</td>
<td>9/19/2018 04:18</td>
<td>9/19/2018 04:18</td>
</tr>
<tr>
<td>Matias</td>
<td>62</td>
<td>8/19/2018 04:18</td>
<td>8/19/2018 04:18</td>
</tr>
<tr>
<td>Daniel</td>
<td>61</td>
<td>7/19/2018 04:18</td>
<td>7/19/2018 04:18</td>
</tr>
<tr>
<td>Andrew</td>
<td>33</td>
<td>6/19/2018 04:18</td>
<td>6/19/2018 04:18</td>
</tr>
<tr>
<td>Dhruv</td>
<td>20</td>
<td>5/19/2018 04:18</td>
<td>5/19/2018 04:18</td>
</tr>
<tr>
<td>Paul</td>
<td>19</td>
<td>4/19/2018 04:18</td>
<td>4/19/2018 04:18</td>
</tr>
<tr>
<td>Hamed</td>
<td>11</td>
<td>3/19/2018 04:18</td>
<td>3/19/2018 04:18</td>
</tr>
<tr>
<td>Hongzhe</td>
<td>10</td>
<td>2/19/2018 04:18</td>
<td>2/19/2018 04:18</td>
</tr>
</tbody>
</table>

#### Last Connect Developers

- **Inactivity Period:** 30 days
- **Last Connect:** 9/19/2018 04:18
- **Most Active Authors:**
  - Sebastian
  - Matias
  - Daniel
  - Andrew
  - Dhruv
  - Paul
  - Hamed
  - Hongzhe

#### Graphs and Statistics

- **Active Developers:** 103
- **Total Commits:** 919
- **Most Active Commits:**
  - Sebastian
  - Matias
  - Daniel
  - Andrew
  - Dhruv
  - Paul
  - Hamed
  - Hongzhe

---

**Bitergia**

Open Containers Initiative Demography panel
Gender-diversity Analysis of the Linux Kernel Technical Contributions

Women in OpenStack report for WOO Meeting

Gender-diversity analysis of technical contributions in the Hadoop Ecosystem
Open NFV code review stats

Development cycle analysis

Idea, request → Iterations → Testing → Deployment
Some reviewers are more equal than others

Source: Some developers are more equal than others (Bitergia’s blog, 2015)

Neutrality?

Source: Understanding How Companies Interact with Free Software Communities (IEEE, 2013)
Data driven community management
That’s not all...
The Cauldron

GitHub organizations analysis

Up to 5 GitHub organizations per user

Latest 30 active repos per organization analysis

Snapshot (no data update)

Kibana 5.x based dashboards

FREE

http://cauldron.io
Are you a community manager consultant?

Are you working as developer advocate for other companies?

We are looking for partners!!

Ask us about our partnership program: info@bitergia.com
Data Science for Community Managers

J. Manrique López de la Fuente
@jsmanrique
jsmanrique at bitergia dot com
https://speakerdeck.com/bitergia