Geo-Locating the Knowledge Transfer in Stack Overflow

Dennis Schenk, Mircea Lungu
Software Composition Group
University of Bern
Switzerland

***Abstract*—Stack Overflow can be seen as an information market for software engineering knowledge in which the goods that are exchanged are answers to questions and the rewards are score points and badges that contribute to a users reputation. By analyzing the transactions in Stack Overflow we can get a glimpse of the way in which the different geographical regions in the world contribute to the knowledge market represented by the website. In this paper we aggregate the knowledge transfer from the level of the users to the level of geographical regions and learn that Europe and North America are the principal and virtually equal contributors; Asia comes as a distant third, mainly represented by India; and Oceania contributes less than Asia but more than South America and Africa together.**

***Index Terms*—Stack Overflow, visualization, case study**

I. Introduction

Stack Overflow (SO) is a Q&A web site for programming knowledge, which can be seen as an open market for information in which the buyers are looking for knowledge (i.e., answers to their questions) and the sellers are recompensed with score points and badges contributing towards their reputation and power[1] on the market. SO can be viewed as a collective effort to solve common problems of those who participate [12] with motivational drivers very similar to those to be found in scientific and OSS communities [8]. Badges used on SO have also been shown to increase and steer users behaviour [2], [7].

SO provides large part of its dataset publicly[2]. As a result, recent years have seen many researchers using it as a case study for a variety of goals: large graph visualization [1], IDE improvements based on the information retrieved from SO [4], studying the kind of topics discussed on the website [5], studying the contribution of participants of different ages [10], etc.

In this paper we look at the state of the SO market from a geographical perspective. We are interested in getting a big picture representing SO, whitout trying to interpret and conclude too much from such a specific snapshot, which would go beyond the scope of this paper. To achieve such an overview we aggregate information from the individual user level to higher geographical levels. To be able to do such an analysis we need to properly geolocate the users of the site and then assemble the information to the country and continent levels. This approach is similar to the one of Bird and Nagappan who localize the contributions of the participants to Eclipse and Mozzila projects [6].

The structure of this paper is the following: in Section II we present the details of the data set we use and the geo-location approach and in Sections III and IV we present some of the observations that follow from our analysis. In Section V we discuss some of the threats to the validity of the analysis.

II. The Data

To perform our analysis we started from the curated dataset provided by Bacchelli [3].

SO allows users to create a profile and optionally specify a location. However, since this information is optional we expect that not all the users will provide it. Many users might not provide their geographical information due to privacy reasons. On top of that, there are many questions and answers to which anonymous or deleted users are attached. Our first questions are thus: *Are those users that provide location information a significant percent of the total user population? How much do they contribute to the total SO knowledge economy?*

By analyzing the data we discover that less than 20% of users (250K users) provide their location information. However, many users create accounts that they never use so maybe many of these users also do not bother to provide detailed information in their profile.

In our analysis we consider the *wealth of an entity* in the market as being their SO reputation. This wealth can be aggregated from the individual to the geo-political entity, as we will see later. We now look at how much of the wealth is accumulated by the users that are geo-located. We sum up the reputation points for the geolocated users and we discover that 75% of the reputation points are distributed to this minority. For the remainder of this paper we only consider this subset of users. We also only look at question and answer relationships that both have geo-locatable owners.

Observation 1. *The minority of 20% of the users which provide their geo-location information in SO collect 75% of the wealth in the market.*

We used the Yahoo Geolocation API and enriched the original dataset with the additional geographical information per user. One of the most interesting observations is that we have in this way discovered users from 198 countries, which is more or less the accepted number of countries in the world.

[1] Different actions on the site require minimum reputation levels

[2] http://www.clearbits.net/creators/146-stack-exchange-data-dump

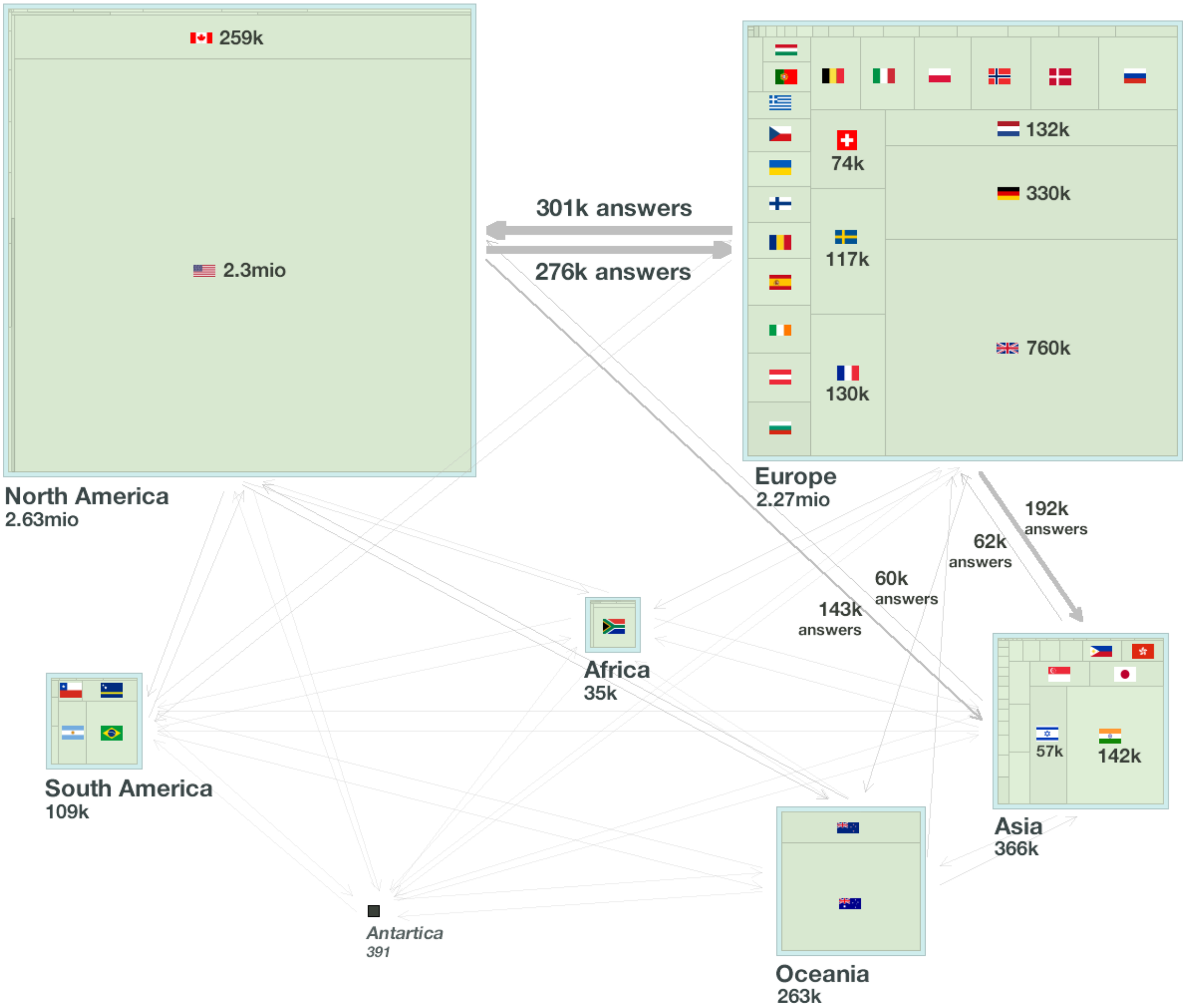

Fig. 1. Information flow in SO aggregated to the continent level. Size and numbers for continents and countries represent the accumulated scores received by their users for answers given.

III. The Big Picture

The information transactions on SO can be seen as a *knowledge flow graph* with users as nodes and information exchanges as edges starting from the user (or users) providing answers and ending at the user receiving the answers. Since there are millions of such transactions that we want to aggregate at the geo-political level we use the automated tool support provided by our tool called Quicksilver[3].

Quicksilver is part of the Moose analysis framework [11] and is an evolution of Softwarenaut [9] our software analysis tool aimed at supporting interactive visualization and exploration of software. In Quicksilver we are porting the lessons learned in Softwarenaut towards a domain independent hierarchical graph analyzer. The hierarchical graph in our case can be built by aggregating the knowledge flow graph information to the level of countries and continents. The user received answer scores can also be aggregated to the geographical level.

Figure 1 presents a visualization obtained with Quicksilver on the information flows between the countries in the world through SO. The tool is interactive and allows the user to drill down in the graph and search for certain nodes. To mitigate the lack of interaction in the medium of this article we have annotated the figure with relevant information where needed. We summarize the construction principles of the visualization in Quicksilver:

- The leaf elements of the hierarchical graph are users which have provided valid locations in their profiles.
- The relationships between users represent answers, starting from the provider and ending at the receiver. The receiver is the one who has posted the corresponding question.
- The users are aggregated up into geo-political units; this aggregation propagates the score points given for

[3] See http://scg.unibe.ch/research/quicksilver

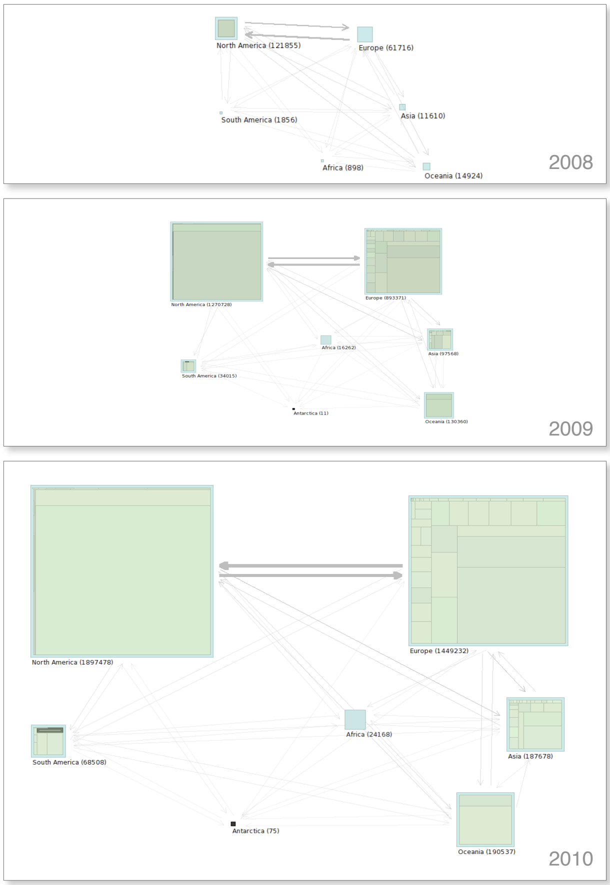

Fig. 2. Three moments in the history of SO: (top) at the end of the private beta in mid September 2008; (middle) at the end of 2009, and (bottom) at the end of 2010. One sees that the overall configuration of the distribution of reputation and flow of information in SO shows little variance over the years.

the answers of the users to the level of countries, and continents.

- The area of the geo-political units is proportional to the aggregated answer score points; the larger an area, the more score points it represents.
- The arrows represent information flow, answers, aggregated form the user level; the thicker and darker an arrow, the more answers it represents.
- The darker the color of a surface, the higher the average received answer scores of the users in that region.

The first thing that hits the eye is that the SO conversation is global although most of the information exchange happens between Europe and North America, and most of the score points are also cumulated in these two continents. Both continents have similar scores and exchange similar amounts of information.

The United States and Canada are the only countries visible in North America. Although the countries of Central America are also included in the graph their aggregated answer score is barely visible.

In Europe, although in principle all Union countries are visible and contribute, the contributions for Germany and UK add up to about as much as all the other European countries combined. However, the reason for the UK contributing more to the information exchange on the website might be related to the fact that the language of the website is English.

Observation 2. *The information exchange in SO is global. However North America and Europe are the main contributors to the knowledge base on the website.*

India has the most answer score accumulated in Asia, followed by Israel and Japan. This particular ranking might reflect the affinity of these countries to the English language although this would remain to be verified. Overall however, Asia is a disproportionately large importer of knowledge with respect to its exports. One interesting observation in Asia is the higher average reputation of the users in Israel.

Observation 3. *Asia, Oceania and South America contribute much less than Europe and America to the global discussion happening in SO.*

We were curious about the participation of Antarctica. The idea of people coding in the loneliness of the continent and using SO intensely (they have the highest questions and answers numbers per participant on average) was enticing. After investigating the nine users who had put Antarctica as their location we concluded that most had probably provided misleading location information[4]

IV. Evolution of the market

One aspect that the previous analysis does not answer is the evolution of the SO over the years. How did it spread, was it a global phenomenon from the beginning? Figure 2 visualizes the evolution of the user activity from the beginning of the website.

The first image on top represents the interval between Jul 2008 to Sept 2008 the time in which the site was in a private beta stage. Even in this stage the global distribution of participants and flow of information as we have seen in the big picture is already, with some variance, established and it will not change much over the years.

Observation 4. *The overall distribution and of flow of information in the market stayed with little variance the same since SO went public, it grew and spread evenly.*

When the site went public it had about 5k geo-located participants. The number of participants almost tripled until the end of 2008. At the end of 2010 there were already 80k participants active and at end of August 2012 there were 250k.

Oceania has a total accumulated answer score of 130k at the end 2009 while Asia had 97k. An interesting development is how Asia grew bigger and surpassed Oceania in the following years. While Oceania had more than double the number of

[4] We tried to contact all the users and two wrote back to us: One user is located in the USA, the other in Asia.

answers per participants than Asia at that time, Asia had more questions per participant. This trend then continued until the end of 2012: Although Asia had only a fourth of the answer scores per participant compared to Oceania, it had more than five times the number of participants than Oceania and thus surpassed it in total answer score (366k vs. 263k).

V. Discussion

For simplicity in this paper we often referred to a country importing or exporting information. However, one must not forget the context in which these terms are used. All the assertions in this paper must be understood to hold only in the context of the SO dataset, and particularly only that part of the data that involves users that have provided a location in their profile.

User Location

Although these users represent only a fraction, a little less than 20% of the total SO user population, this threat is alleviated by the fact, that they accumulate around 75% of all answer scores in the SO knowledge economy.

The user location is the one specified at the moment when the dataset was created, so if the user has moved recently, the previous answers that he gave, while he was living somewhere else, are wrongly associated with his latest location.

We have no way of detecting users that have declared false locations so we assume that not too many of them will provide a false information, especially since publishing the location of a user is not a privacy threat.

Locations can be ambiguous, e.g. there is a Kingston in Jamaica, Canada and the USA. We decided to just use the first one returned by the API.

User Language

SO is a platform where knowledge is traded in English, so the countries that have English as their native language might have an advantage. Although software development is an activity that happens naturally in English observing Germany which has half the reputation of UK although a slightly higher population seems to support the concerns regarding this threat. It would be interesting to study whether users from countries with less affinity to English use alternative platforms although this is unlikely[5].

Traffic Data

We correlate the results presented in this paper with information about the website traffic. Indeed, many users that do not ask or answer questions are nevertheless using the website to find information. In a post from 2011, Joel Spolsky, one of the founders of SO, reports on the demographics of SO visits normalized to the number of users in a country.[6] The top seven countries at the time of his analysis were: Sweeden, Singapore, Finland, Denmark, Israel, Switzerland, and the Netherlands. Note that this top is normalized to the population, so what it tells is that a higher percentage of those countries populations might be programmers.

VI. Conclusion

In this paper we have appraised the state of the knowledge economy in SO by aggregating the individual knowledge transactions to the geographical level. We have discovered that the users that contribute the large majority of the knowledge and collect the large majority of reputation care to provide their own location information. We have observed that Europe and the United States are the strongest contributors in a discourse that involves all the countries in the world. We have observed that Asia is a strong importer of information. Finally by analyzing the evolution in time of the geo-located answers we learned that SO started as a global phenomenon right from the beginning.

References

[1] L. Akoglu, D. H. Chau, U. Kang, D. Koutra, and C. Faloutsos. Opavion: mining and visualization in large graphs. In K. S. Candan, Y. Chen, R. T. Snodgrass, L. Gravano, and A. Fuxman, editors, *SIGMOD Conference*, pages 717–720. ACM, 2012.

[2] A. Anderson, D. Huttenlocher, J. Kleinberg, and J. Leskovec. Steering user behavior with badges. In *Proceedings of the 22nd international conference on World Wide Web*, WWW '13, pages 95–106, Republic and Canton of Geneva, Switzerland, 2013. International World Wide Web Conferences Steering Committee.

[3] A. Bacchelli. Mining challenge 2013: Stack overflow. In *The 10th Working Conference on Mining Software Repositories*, page to appear, 2013.

[4] A. Bacchelli, L. Ponzanelli, and M. Lanza. Harnessing Stack Overflow for the IDE. In *Recommendation Systems for Software Engineering (RSSE), 2012 Third International Workshop on*, pages 26–30. IEEE, June 2012.

[5] A. Barua, S. W. Thomas, and A. E. Hassan. What are developers talking about? an analysis of topics and trends in stack overflow. *Empirical Software Engineering*, page To appear, 2012.

[6] C. Bird and N. Nagappan. Who? where? what? examining distributed development in two large open source projects. In *Mining Software Repositories (MSR), 2012 9th IEEE Working Conference on*, pages 237–246, 2012.

[7] S. Grant and B. Betts. Encouraging user behaviour with achievements: an empirical study. In *Proceedings of the 10th Working Conference on Mining Software Repositories*, MSR '13, pages 65–68, Piscataway, NJ, USA, 2013. IEEE Press.

[8] G. Hertel, S. Niedner, and S. Herrmann. Motivation of software developers in open source projects: an internet-based survey of contributors to the linux kernel. *Research Policy*, 32(7):1159 – 1177, 2003. <ce:title>Open Source Software Development</ce:title>.

[9] M. Lungu, M. Lanza, and O. Nierstrasz. Evolutionary and collaborative software architecture recovery with Softwarenaut. *Science of Computer Programming (SCP)*, 2012.

[10] P. Morrison and E. Murphy-Hill. Is programming knowledge related to age? an exploration of stack overflow. In *Proceedings of the 10th Working Conference on Mining Software Repositories*, MSR '13, pages 69–72, Piscataway, NJ, USA, 2013. IEEE Press.

[11] O. Nierstrasz, S. Ducasse, and T. Gîrba. The story of Moose: an agile reengineering environment. In *Proceedings of the European Software Engineering Conference (ESEC/FSE'05)*, pages 1–10, New York, NY, USA, Sept. 2005. ACM Press. Invited paper.

[12] E. S. Raymond. *The Cathedral and the Bazaar*. O'Reilly & Associates, Inc., Sebastopol, CA, USA, 1st edition, 1999.

[5] In a discussion on a German forum (http://goo.gl/DMLcR) a user searching for a German alternative to SO is directed to a small localized SO clone while advised to stick to English in programming matters

[6] http://blog.stackoverflow.com/2011/04/stack-overflow-around-the-world/