World Wide Web Journal manuscript No.
(will be inserted by the editor)

A Time Decoupling Approach for Studying Forum Dynamics

Supplementary Information

Andrey Kan · Jeffrey Chan · Conor Hayes · Bernie Hogan · James Bailey · Christopher Leckie

the date of receipt and acceptance should be inserted later

1 *SAP* and *TiddlyWiki* Datasets

The *SAP Community Network (SCN)* is a professional social network hosted by *SAP*. It has been set-up in order to provide trusted connections to a large community of *SAP* customers, partners, employees and experts, and by this to offer access to knowledge, insight and rich content about *SAP* solutions and services. We have collected data from 5 (randomly chosen) of their technical support forums (http://forums.sdn.sap.com/). We randomly select 30% of users in the forums. The dataset span is from 2006-01-01 till 2011-01-26 (5 years).

TiddlyWiki is an open source project, and it has several Internet forums dedicated to the discussion of project related issues. Technically, these forums are organized as *Google Groups*, but they have the same structure as typical Internet forums: users, posts, replies, and discussion threads. We collected data from three *TiddlyWiki* groups (http://groups.google.com/group/tiddlyweb/about,
http://groups.google.com/group/tiddlywikidev/about,
http://groups.google.com/group/tiddlywiki/about). We randomly select 30% of users in the forums. The dataset span is from 2005-06-15 till 2011-02-09 (5 years 7 months).

A. Kan (corresponding), J. Bailey, C. Leckie
NICTA Victoria Research Laboratory, Department of Computer Science and Software Engineering, The University of Melbourne, Australia
Tel.: +614 20 768 752
E-mail: akan@csse.unimelb.edu.au

J. Chan, C. Hayes
Digital Enterprise Research Institute, National University of Ireland, Galway, Ireland

J. Chan
Department of Computer Science and Software Engineering, The University of Melbourne, Australia

B. Hogan
Oxford Internet Institute, University of Oxford, United Kingdom

Note that in *Boards.ie*, the path segment $(0,0)$ – $(1,0)$ corresponds to the first user post since user registration, but in *SCN* and *TiddlyWiki* forums the path segment $(0,0)$ – $(1,0)$ corresponds to the first user post since the beginning of the time span of the corresponding dataset. This is not necessarily the first user's post since registration. The registration times were not available in these datasets. However, after a manual inspection of randomly selected users in the corresponding websites, it appears that for each user the registration time is after T_0, and there is no evidence that the path definitions for *Boards.ie*, *SCN*, and *TiddlyWiki* datasets differ.

SAP and *TiddlyWiki* datasets are visualized in Figures 1, 2, and 3. The datasets' statistics are presented in Table 1. In all datasets we observe consistency in user communication and dead zones, as reported in the main text of the paper.

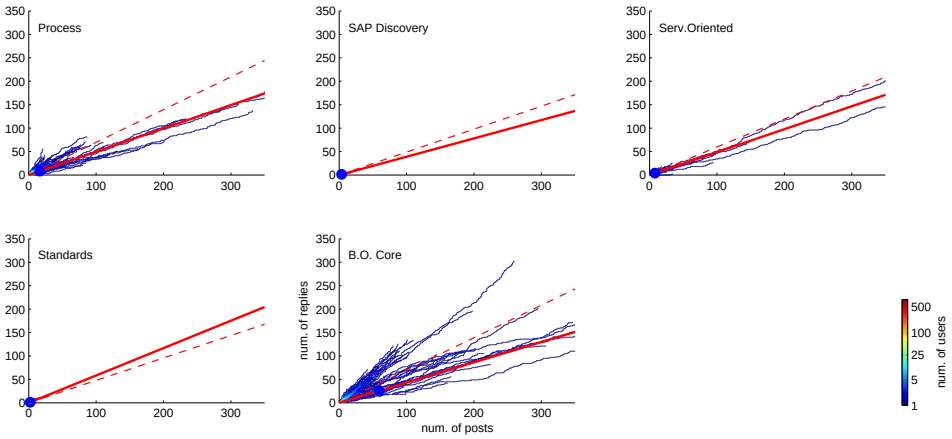

Fig. 1 Five *SCN* forums. Blue dot shows the mean length of user paths; solid red line is the straight line fitted to the forum as a whole; dashed red line reflects the replies to posts ratio (baseline).

Overall, we note that the maximum slope of *SCN* forums is less than the minimum slope for *TiddlyWiki* and *Boards.ie* forums. The selected *SCN* forums represent question and answer type forums in which, once an appropriate answer is received, there is no further need to continue the discussion. On the other hand, some of the forums from *Boards.ie* are also question and answer type forums, yet they have a higher slope. The investigation of this observation is left for future work.

Further, we note that slopes for *TiddlyWiki* forums tend to be lower than slopes of *Boards.ie* forums. A possible reason is that the active *TiddlyWiki* users are passionate about the *TiddlyWiki* project, and this motivates them to provide technical support in forums. We hypothesize that reciprocity of communication in *TiddlyWiki* forums is not that important as in *Boards.ie* forums.

2 Effects of Sampling in *Ancestry.com* Dataset

Ancestry.com is a large genealogy portal, where people can create their family trees, browse historical records or search for relatives. Forum board is an important component of this portal that claims to be "the world's largest online genealogy community

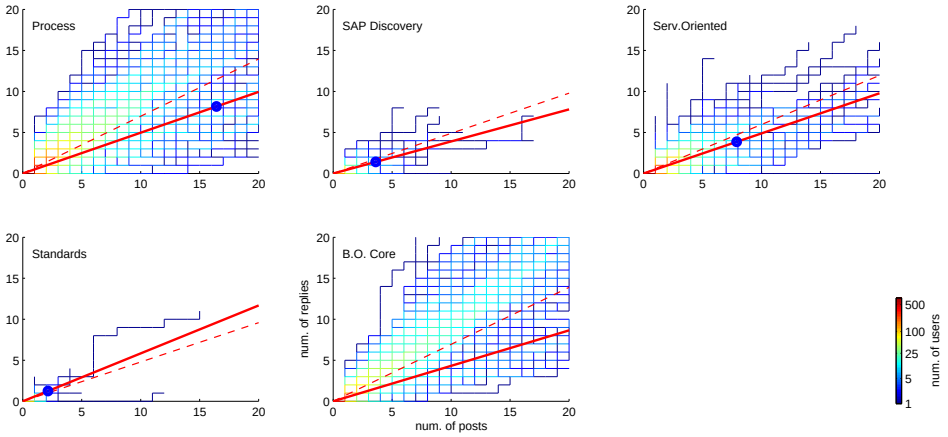

Fig. 2 Five *SCN* forums (close-up view). Blue dot shows the mean length of user paths; solid red line is the straight line fitted to the forum as a whole; dashed red line reflects the replies to posts ratio (baseline).

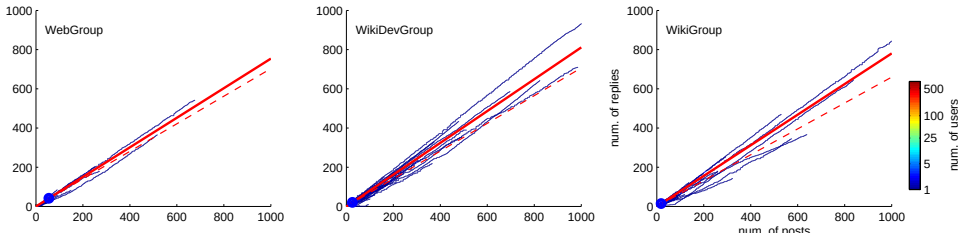

Fig. 3 Three *TiddlyWiki* forums. Blue dot shows the mean length of user paths; solid red line is the straight line fitted to the forum as a whole; dashed red line reflects the replies to posts ratio (baseline).

with over 17 Million posts on more than 161,000 boards."[1] Researchers from *Carnegie Mellon University* in collaboration with *Ancestry.com Inc* have made a full snapshot of the forum data publicly available [1].

The dataset consists of conversation threads, and each thread belongs to one of three categories: "localities", "surnames", and "topics". The first two categories comprise 96% of all messages. Furthermore, the "localities" category is divided into subcategories such as "North America" and "Middle East", whereas the "surnames" category has a flat structure. It consists of forums such that each forum is dedicated to a particular surname (e.g., "Abigail", "Hoffman"). We decided to look at two randomly selected locality subcategories, because the number of threads in such subcategories is large, compared to the number of threads dedicated to a particular surname.

The two randomly selected subcategories are "Central Europe" and "Oceania". Each subcategory is a set of conversation threads related to the corresponding geographical region. We refer to each subcategory as a forum. The original dataset spans a range from December 1995 till July 2010, but we study a period of 4.5 years (2006 – mid-2010). In our study we only consider users that started their activity since January

[1] Source: http://boards.ancestry.com.au/, retrieved on the 10-th of November, 2011

Table 1 Statistics for selected *SCN* and *TiddlyWiki* forums. Number of posts is the total number of posts made by all users from the sample, and number of replies is the total number of replies received by all users from the sample. Length, Slope, Base, and Spread are forum features that are defined in the main text.

Forum	Num. of users	Num. of posts	Num. of replies	Length	Slope	Base	Spread
Process Integration	778	8,406	5,867	18.35	0.5	0.7	8.7
SAP Discovery System for Enterprise SOA	124	323	158	3.88	0.4	0.5	1.07
Service-Oriented Architecture	658	3,620	2,162	8.79	0.5	0.6	12.46
Standards	71	119	57	2.48	0.58	0.48	1.47
SAP Business One Core	300	11,446	7,949	64.65	0.43	0.7	31.78
TiddlyWeb	78	3166	2227	69.14	0.75	0.7	17.81
TiddlyWikiDev	759	14703	10389	33.06	0.81	0.71	31.63
TiddlyWiki	901	12633	8337	23.27	0.78	0.66	29.91

2006, because there was a large reorganization of the forum website that resulted in significant fluctuations of major forum statistics. The fluctuations are observed roughly until 2005 [1].

Table 2 shows statistics for each forum, and Figure 4 shows our visualization for the forums. Note that the numbers and visualizations are given for all users that started their activity since January 2006.

Table 2 Statistics for two selected *Ancestry.com* forums. Numbers are calculated taking into consideration all users that started their activity since 2006.

Forum	Num. of users	Num. of posts	Num. of replies	Length	Base	Offset	Spread
Central Europe	15,338	41,089	24,710	4.290	0.601	-0.170	9.368
Oceania	10,405	43,650	28,725	6.956	0.658	-0.055	11.687

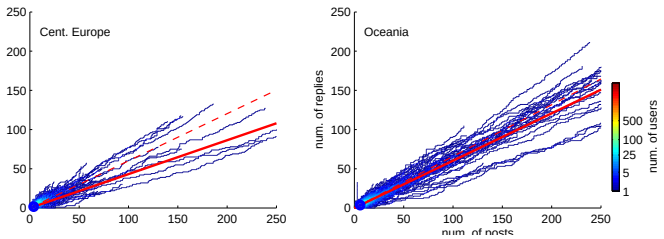

Fig. 4 Visualization of two selected *Ancestry.com* forums. Blue dot shows the mean length of user paths; solid red line is the straight line fitted to the forum as a whole; dashed red line reflects the replies to posts ratio (baseline).

We study effects of sampling users on the estimation of forum features on the two *Ancestry.com* forums because these two forums have the largest number of users across

our datasets. We produced 40 random 30% samples of users from each of the forums, and calculated forum features for each of the samples. Table 3 shows that the feature values for the original dataset always fall into the interval formed by the mean plus and minus one standard deviation of the values estimated from the samples.

Table 3 Forum features for two selected *Ancestry.com* forums estimated from samples of users and from all users in the dataset. For the values that were estimated from the samples the range shows the mean plus and minus one standard deviation.

	Central Europe (samples)	**Central Europe (all users)**	**Oceania (samples)**	**Oceania (all users)**
Length	[3.712; 5.280]	4.290	[6.469; 7.409]	6.956
Base	[0.565; 0.627]	0.601	[0.648; 0.666]	0.658
Offset	[-0.191; -0.043]	-0.170	[-0.093; -0.022]	-0.055
Spread	[3.163; 13.269]	9.368	[7.046; 13.364]	11.687

3 Additional Information for *Boards.ie* Dataset

In Section 6 of the main text, we present visualizations of the nine *Boards.ie* forums, and the visualizations show all prominent forum properties that we discuss in this paper, for example, the slopes and the mean lengths of paths. However, in the main text we show a close-up view of forums to make individual user paths distinguishable. Here, Figure 5 shows a global view of the nine *Boards.ie* forums. From the global view one can observe the same temporal patterns as discussed in the main text: (i) user paths tend to follow a straight line; (ii) user paths tend to to cross certain areas, e.g., the areas that lie close to the x or y axis.

We observe an exceptionally long user path in the gigs forum. The user with ID 60385 has a path length of 1342 whereas a mean length for the gigs forum is 11. This user has a large number of points in his/her path, and has a much higher influence on the slope of the gigs forum than other users. Therefore we truncate the path of this user to the mean length of 11. This is the only case when we truncate an empirically observed user path. As far as we can tell from the dataset, this user is neither a bot, nor a spammer. Presumably, this user is a genuine fan of musical performances.

References

1. Elsas, J.: The Ancestry.com Forum Dataset. Technical Report CMU-LTI-017 (2011)

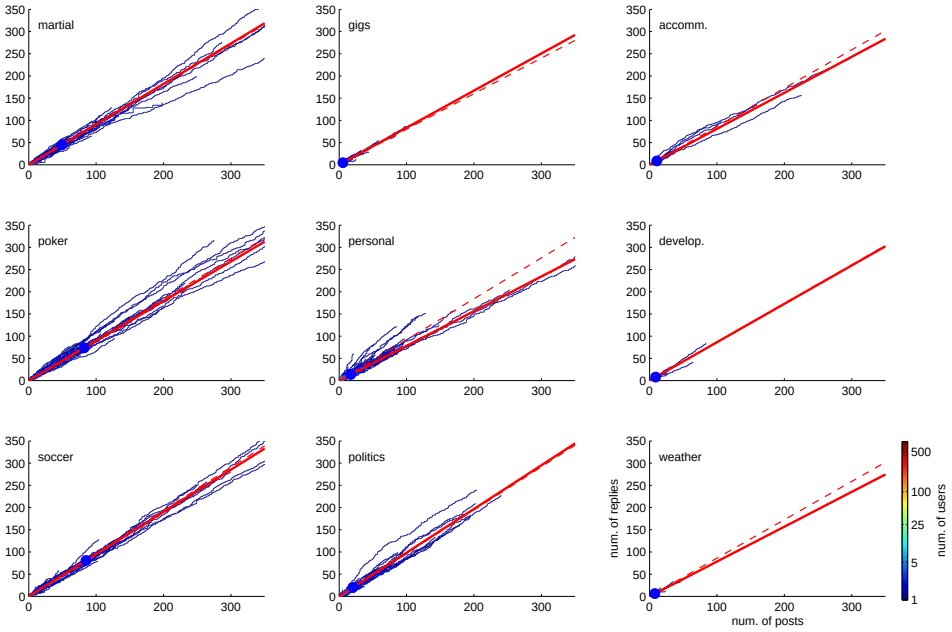

Fig. 5 A view on a larger scale of the nine *Boards.ie* forums. Blue dot shows the mean length of user paths; solid red line is the straight line fitted to the forum as a whole; dashed red line reflects the replies to posts ratio (baseline).