Livrable L2.2-4

"Report on Facebook labelled personal networks"

(M36)

PROJET ALGOPOL

PROGRAMME ANR CONTINT 2012

Partenaire responsable:	LIAFA
Autres participants:	Orange Labs, Linkfluence

Ce livrable présente le travail de catégorisation des métadonnées de Facebook collectées par l'application Algopol. Il est constitué :
– d'un article non publié décrivant la méthode de recodage ;
– du code en python effectuant la catégorisation, disponible publiquement à l'adresse suivante : https://github.com/Algopol/guessed_types

Searching Interaction and Expression in Facebook Metadata

Irène Bastard, Baptiste Fontaine, Stéphane Raux, Dominique Cardon, Christophe Prieur

Abstract

To address sociological questions about the content of interaction ties on Facebook, we faced the problem of the high versatility of Facebook metadata.

The aim of this paper is to share our experience (and the produced code) on what has been mainly a work of reverse-engineering Facebook metadata, so that research communities will use and improve it in subsequent empirical analyses on Facebook.

Keywords

facebook; reverse engineering; metadata; online interaction; social network services

To explain Facebook to a hypothetical non-user of the Web, a simple comparison with the bedroom wall of a teenager is enough: one finds there music star posters, quotations and punchlines on post-its, photos and postcards from friends, birthday wishes and own personal notes and reminders. Now to deal with analysis, Facebook metadata sometimes look such a mess as a teenager bedroom… Among Algopol[1], a project gathering sociologists and computer scientists in order to study social and algorithmical recommendations on the Web, we study Facebook, as a place where those considerations seem critical: media, profiles, groups, mix affective and technical recommendations. Whether one can publish news, opinions, feelings or photos, the public, addressed and concerned, is crucial to understand the context of expression. Traditional works in sociology have been using surveys (Ellison et al., 2014; Steinfield et al., 2008; Livingstone et al., 2012) or ethnographies (boyd, 2008; Ito et al., 2010) to study online social activities, questioning profile's construction, personal expression and social development. Now a natural way to study users' behaviors in social network services is to analyze logs of actual interaction with the platforms, which requires mining of large datasets. This trend has become common, to characterize user behavior (Benevenuto et al, 2009, Schneider et al, 2009), or to estimate tie strength of pairs of individuals (Gilbert & Karahalios, 2009; Viswanath et al, 2009; Xiang et al, 2010).

In a survey based on a Facebook application, and with the help of CSA, a poll institute, we have gathered a sample of 877 respondents, representative of French Facebook users, from which we have collected the Facebook profile informations, the ties between their friends, and all that had been posted on the respondents' walls (which includes information about their activity on other users' walls). Before analyzing this dataset to question the types of interaction between users and the platform, among users, the way they share content, or the structure of their networks, we faced the problem of decoding Facebook's metadata. Filling the gap between what metadata is supposed to mean and what it actually means, and translating it into an accurate information, was far from being an easy task. What seems interesting in this dirty work is that it forces to look at Facebook with special glasses which make a cut between expression and interaction. Facebook is not a place where one cares only about one's own self, most of the activities on the website are interactional: tagging someone in a photo, *lol*-ing to a friend's joke, greeting a relative on their wall for their birthday, etc. All of those actions embed user expression

1 Granted by the French national research agency (ANR-12-CORD-018)

in an interaction, and metadata about the activity has to be put into its social context. The aim of this paper is to share our experience (and the produced code) on what has been mainly a work of reverse-engineering of Facebook metadata, so that research communities will use and improve it in subsequent empirical analyses on Facebook.
After giving some technical and empirical keys to Facebook metadata about wall posts, we will describe an analytical framework to consider activity through five features, with some clues on how we managed to compute them from metadata. We will then give some results on the distribution of each feature in our sample, and give the link to our Python code which characterizes each activity item provided by the Facebook API, among 92 activity types.

From Facebook metadata to Algopol's "guessed_type"

For a given respondent, that we will call Ego, the Facebook API let us collect information, with Ego's consent, on their profile, their friends' profiles, and all that was posted on Ego's wall[2]. In this paper, we consider only data about posts.
The Algopol application uses the API to produce one JSON file per Ego, listing all activities that were posted on their wall since the account's creation[3]. In this file, all entries, called *statuses*[4], contain multiple fields: "*from_id*" identifies the activity's author, "*message*" contains the text of the post if it is a status, "*created*" is the creation date. We initially looked at the "*type*" field, supposed to inform on the type of the post. These fields are particularly unstable and inconsistent: unstable because features are added and removed along time, for instance, the "music" type appeared at some point then disappeared later; they are also inconsistent because the same activities performed on a desktop computer or on a mobile will not have the same type, some awkward examples will be given in the next section. Table 1 lists the number of occurrences of all the identified status types in our dataset. It represents 32 combinations with amounts from almost 1 million entries to only 2. Facebook metadata, and particularly the "*type*" field, does not allow us to describe the users' activity, neither self-expression nor the interactional context. The API documentation does not help much here because it fails at precisely describe statuses for some categories. When Ego adds a new photo, Facebook does not tell us in a clear way if they are changing their profile picture, cover photo or just adding it to their album. The documentation is also unclear about profile updates such as Ego changing their education information or current city, as well as their relationship status updates.
We thus decided to compute a new field we called "*guessed_type*" to more accurately describe Ego's activity. Our work led us to a list of 92 guessed_types, exploring accounts' activity, sometimes one status at a time, and combining information from multiple metadata fields. We exploited the "*story*" field: it originally describes a user activity as a text, such that "Ego commented on the status of…" or "Ego is now friends with Alter". This field is the place where

2 Some research open the Facebook box from other entry points, such as political pages, or messages with specific content (Ellison *et al.*, 2013). A specificity of our approach is to start from users data to look to content, inversing a more common logic which looks at web contents and next searching audience and sociodemographic analysis of users.
3 We obviously have neither access to private messages, nor to the "remorse", the deleted Facebook status (Das, Kramer, 2013).
4 Note that all activities reported in the file are called statuses, not only the common « status » post, consisting of a text posted by a user on their wall.

Facebook automatically creates a meta-text of Ego's expression. We used it to gather more information about the type of a status.
Those guessed_types are devised to be mutually exclusive, and to cover all kinds of expression and interaction on Facebook. We detected 33 different guessed_types to refine the common API field "status" in order to distinguish between various types of photos, videos, text statuses, links or events. The challenge here is to find systematic rules to describe an online activity in its social context, despite the unstability of Facebook platform features, metadata, and large variety of uses (for instance, people use several artifacts to tag their friends in photos, either by directly using Facebook's tagging feature, or by mentioning them in the description).

Table 1 – Combinations of status types

985511	status
971598	link, app_created_story
128970	photo, shared_story
112212	status, mobile_status_update
108516	link
98798	status, wall_post
93937	link, shared_story
83870	status, approved_friend
51056	video, shared_story
49579	photo, added_photos
27632	photo
16821	status, app_created_story
13620	photo, tagged_in_photo
10138	swf, app_created_story
8781	question
6568	video
5648	link, approved_friend
2159	video, added_video
1680	checkin
1555	photo, mobile_status_update
345	link, created_note
301	link, created_event
287	video, app_created_story
171	offer
118	status, shared_story
47	link, created_group
39	music, shared_story
19	offer, shared_story
7	music, app_created_story
6	music
2	swf
2	photo, app_created_story

Describe each activity with five features

We compute a vector on each activity reported by the Facebook API. This vector is supposed to properly place Ego expression in its interactional context through online artifacts due to these five components: "Who?", "What?", "Where?", "With whom?", "Words". Each of these components is built with a combination of Facebook metadata fields, based on explorations, back and forth across specific examples and generic rules.

Who? – This first question cannot be easily answered from the Facebook API despite the "*from*" field of statuses, because but not all statuses come from users. Applications can post messages as Ego on their wall even if the user isn't the author. One can notice this case when applications repeatedly post English messages on the wall of a non-English speaker. The Facebook API provides an "*application*" field which contains the application used for the publication, but this as well is not sufficient as some of them are games posting on the user's behalf while others are media websites the user purposedly shared a story from, mobile platforms or even Facebook's

own features like Links and Pages. To identify if the status is posted by Ego or not, one can first check the application field, the message and the publication's URL to determine if it comes from a game or not, and next compare Ego's id to the *from_id* field. An activity which is not in the context of an application or a game might have been performed by *Ego* or by someone else, that we call *Alter*[5].

What? – Our second feature is "What", which tells us which kind of activity we are looking at. Here again, Facebook provides a designated field called "*type*" but as was mentioned above, it is not sufficient. The status type of an event participation is "*link*", with the URL of the event's page. Likes on pages, links, photos or text publications all have either the type "status" or the type "*link*". Both a change of the profile picture and a photo posted on a friend's wall get the type "*photo*" and nothing more in the field. Most of our tests are made here on the text content of the *story* field. Facebook uses quotes for user-written content so we can check the text outside of them as a kind of unofficial metadata, without the risk of false-positives with user content. Our "*What"* field may be filled with "*status"*, "*photos"* or "*link"*, but some other activities do not use specific content, for instance on friends' approvals we set the "*What*" field to the value "*Facebook*". Note that the metadata do not provide the information whether the object shared is public or not: a photo shared by Ego can for instance come from a football team's official page and be public, or from my friend Ibra and be private.

Where? – The "*story*" field is written with lots of declinations of objects, according to the place they come from: "*picture from Firstname Lastname*" is not the same as "*Photo from ID*" or "*Photo*" … The "*Where*" field is a deconstruction of the Facebook story to establish whether it reports about an activity posted on Ego's wall or on a friend's wall. Comments or statuses on friends walls are present in the dataset only if they appear at the same time on Ego's timeline, in which case the story is "*in the timeline of Firstname Lastname*". The possible values for the field "*Where*" are *Ego's wall*, *Alter's wall*, and *Page* (which includes pages of personalities, applications or brands, but also group pages).

With whom? – This feature is the one which stays the more complex, because it inherits the same undetermination than the "*Who*". In the long history of Facebook features, profiles have been launched before groups and pages, and who knows what is coming. The interactional dimension of a status is established by the mention of someone, in a text, or a tag, in a photo. The fields "*tags*" and "*story_tags*" deal with those designations. Finally, we considered the field "*to*" in order to identify Ego's activity in a page context. It makes the distinction of expression *Alone*, with *Friend*, in a *Page* (which is redundant with the "*where*" field), and with a *quote*.

Words? – The interaction with web platforms mixes words and clicks. In Facebook, a "like" is an expression as well as words. The feature "*Words*" refers to the action of expression. We consider two things in this field: Ego's message if there is one, and the action they performed. Most of the posts, as they are described in the "*story*" field, include a verb, like "participate" for an event or "*add*" for albums. We used them to infer the action associated with the activity.

5 this Alter is not necessarily a friend of Ego, depending on Ego's privacy settings.

Possible values for our field "*Words*" are *words* (status, photos with text, comment), *publish* (link or photo without text, including a profile picture), *like* and *share*.

The 5-components vector is used to describe an activity and its context. Facebook produces interlinked contexts: Ego likes a status of a friend about a photo of a group, or more precisely, we found some posts stories such as "Ego likes what you liked". An interesting work track might be to trace the sequences of statuses, comments and likes: if Ego received a *like* from a friend on a status, will they *like* some content of that friend in the next days?

Results and discussion

This section aims at experimenting the vector on our dataset. Table 2 below presents the volume of activities collected from the 877 respondents of our sample, ranging from 1 to 2.8 millions of lines of activity for each Ego. We have excluded the activity generated by games and applications without Ego's own expression (37% of all activities), and rejected 9 unqualified guessed_types (for instance, some activity presents no story, url, nor message; it could be Deezer listening publications on one's wall, but without any indications we cannot accurately qualify the activity).
Those results establish that activities on Facebook are diverse, for instance we count in our dataset 24% of activities with photos, 17% with links, 36% of likes and 32% of words from Ego. The success of each activity is also various, a photo is, on average, a great factor of likes from friends (especially the profile photo, gathering on average 7.5 likes and 2.5 comments), and posting on a page is one of the most commented activity.

With this empirical confrontation of our framework to the data, some vector combinations in guessed_types are possible and others are not: the vector yields 384 possibilities while we actually count "only" 92 guessed_types ("Alter likes a Photo on Alter's page" cannot appear in our sample, because it is only an Alter's activity without Ego's implication). Even among Ego's activity, commenting on a photo posted by Ego on Alter's wall is seen, in Facebook metadata, as a comment of Ego's photo, because the picture is automatically added to Ego's personal photos, so the corresponding guessed_type is empty. On the other side, Facebook metadata counts 48,305 photos in our dataset, but it is impossible to link those pictures as profile photo update (4,291), or as photo with citation (1,597) or just as added photo with text (18,320) or without text (57,139; which is more than photos in the profile account, showing that activity is sometimes saved in the portfolio and sometimes not).

Table 2

	Guessed_ type	Activity Count	%_Activity	Avg. Like / A	Avg. Comm / A
Total	**92**	**2 722 367**		**0,66**	**0,44**
App / Game	1	1 005 531	37%	0,31	0,07
Total Activity	**91**	**1 716 836**		**1,63**	**1,04**
Unqualified	9	49 323	2%	3,97	2,32
Who					
Ego	65	1 562 190	91%	0,97	0,73
Alter	6	105 323	6%	5,17	2,39
What					
Status	16	554 085	32%	0,56	0,86
Photo	19	407 989	24%	3,49	1,66
Facebook	16	412 264	24%	0,48	0,32
Link	19	292 813	17%	0,71	0,62
Where					
Ego's Wall	53	1 334 264	78%	1,54	0,94
Alter's Wall	10	300 486	18%	0,01	0,00
Page	8	32 763	2%	1,43	1,41
With Whom					
Alone	38	1 116 521	65%	0,85	0,56
Friend	20	509 058	30%	1,64	0,76
Page	8	32 763	2%	1,43	1,41
Quotation	5	9 171	1%	3,65	2,84
Words					
Like	10	619 010	36%	0,01	0,01
Words	30	544 037	32%	1,20	1,21
Publish	20	319 933	19%	2,49	1,11
Share	11	184 533	11%	0,78	0,34

Conclusion and Perspectives

Facebook metadata are the result of many changes in the long history of the platform features, so using them to analyze the activity of users requires a careful combination of the fields provided by the API. We have given here many examples of glitches in the way metadata give information about actual activity, and provided a methodological framework to overcome this, based on a 5-features vector. We would like to claim for an open research on Facebook data, which is why we have made publicly available for the community the Python code we have written to classify activities among the 92 so-called guessed_types[6]. We do not claim that the our 5-features vector is universal, but it is likely to meet many of the needs of other research on

6 https://github.com/Algopol/guessed_types

online interaction, including in other social network services (as LinkedIn), mobile applications (as Snapchat), or other mediated interactions (as phone texts).

Our next research efforts, among the Algopol project, will be directed to the analyze of activity profiles on Facebook, for instance to describe the use of citations by teenagers, and to more precisely explore content practices like URL references. Our application now has over 16,000 users, which provides us with enough respondents to build sub-samples with specific profiles like students or elder people without accuracy problems, usual on small datasets. These activity items will also be studied in conjunction with the social network structure of egos, to take into account the audience of egos' expression, and to characterize the nature of Facebook ties.

The work we have described here raises another interesting sociological question directed on the way data are produced and organized (Garfinkel & Bittner, 1967). The relative disorder of Facebook metadata is likely to be the result of the internal processes of Facebook as an organization, whose main outcome is not datasets, but a web platform with ever-evolving features with a constant balance between many contradictory expectations.

Bibliography

Benevenuto, F., Rodrigues, T., Cha, M., & Almeida, V. (2009). Characterizing user behavior in online social networks. *Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement Conference IMC 09*, 49. doi:10.1145/1644893.1644900

Boyd, D. (2008). *Taken Out of Context*. University of California, Berkeley.

Das, S., & Kramer, A. (2013). Self-Censorship on Facebook. In *AAAI*.

Ellison, N. B., Steinfield, C., & Lampe, C. (2007). The Benefits of Facebook "Friends:" Social Capital and College Students' Use of Online Social Network Sites. *Journal of Computer-Mediated Communication*, *12*(4), 1143–1168.

Ellison, N., Gray, R., & Vitak, J. (2013). Calling All Facebook Friends: Exploring Requests for Help on Facebook. *ICWSM*.

Garfinkel, H., & Bittner, E. (1967). Good organizational reasons for "bad" clinic records. *Englewood Cliffs*.

Gilbert, E., & Karahalios, K. (2009). Predicting Tie Strength With Social Media. In *CHI'09: Proceedings of the 27th annual SIGCHI conference on Human Factors in Computing Systems*. New York, NY: ACM Press.

Ito, M., Baumer, S., Bittanti, M., Boyd, D., Cody, R., Herr-Stephenson, B., … Tripp, L. (2010). *Hanging out, messing around, and geeking out*.

Livingstone, S., Haddon, L., Gorzig, A. & Olaffson, K., Risks and Safety on the Internet: The Perspective of European Children. Full Findings, Londres, LSE, EU Kids Online, 2010.

Schneider, F., Feldmann, A., Krishnamurthy, B., & Willinger, W. (2009). Understanding online social network usage from a network perspective. *Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement Conference IMC 09*, 35. doi:10.1145/1644893.1644899

Steinfield, C., Ellison, N. B., & Lampe, C. (2008). Social capital, self-esteem, and use of online social network sites: A longitudinal analysis. *Journal of Applied Developmental Psychology*, *29*(6), 434–445.

Viswanath, B., Mislove, A., Cha, M., & Gummadi, K. P. (2009). On the Evolution of User Interaction in Facebook. *Proceedings of the 2nd ACM Workshop on Online Social Networks - WOSN '09*, 37. doi:10.1145/1592665.1592675
Xiang, R., Neville, J., & Rogati, M. (2010). Modeling Relationship Strength in Online Social Networks. In *WWW 2010* (pp. 981–990). doi:10.1145/1772690.1772790