Thematic and Social Indicators for Flickr Groups

Christophe Prieur[1], Nicolas Pissard[2], Jean-Samuel Beuscart[2], Dominique Cardon[2], Pascal Pons[1]

[1] Liafa, University Paris-Diderot
case 7014, blabla, 75205 Paris Cedex 13,
France

[2] Orange Labs
TECH/SENSE, 38 rue du Gal Leclerc,
92131 Issy-les-Moulineaux Cedex

prieur@liafa.jussieu.fr, npissard@yahoo.com,
jeansamuel.beuscart@orange-ftgroup.com,
domi.cardon@orange-ftgroup.com, pascal.pons@gmail.com

Abstract

We study here Flickr groups in order to see whether they are actual communities or rather essentially thematic clusters. We describe a methodological framework for the analysis of networks of group members, measuring social density and tag dispersion among groups and give some results on a sample of 450 groups having around 500 members each.

Introduction

As a result of the spread of self-production tools, Web 2.0 services enable cooperation between Internet users as a side effect of their individual publication activities. The 'strength of weak cooperation' (Aguiton, Cardon, 2007) lies in the fact that it is not necessary for individuals to have a cooperative plan of action or an altruistic concern beforehand. They discover cooperative opportunities simply by making their individual productions public. A typical example of this process is Flickr: not only a website for photo publication, it also provides tools that enable coordination. Our goal here is to sketch a way to study Flickr groups as a key element of this weak cooperation.

Our experiments are done on a sample of an extensive database we have collected from the Flickr website and whose detailed figures and analysis will be published soon (Prieur *et al*, 2008).

Flickr Groups: Thematic *and* Social Tool

Flickr(.com) is a website that enables users to upload photos, index them with freely chosen keywords called *tags* (cat, paris, etc.) and post them to thematic user-created *groups* (*"Cats rule"*, *"People in the street"*, etc.) They can also put *comments* on other users' photos, mark them as their *favorites* and mark these users as their *contacts*.

Among the site's functionalities, tags, contacts and groups are the three giving direct access to photos. The first two have very distinct functions: tags are essentially used for indexing — a photo with the tag cat will appear in global searches made on this tag. As for contacts, they are the core material of the social media — Flickr shows you the recent photos of your contacts with the idea that people don't only want to see photos of something but also someone's photos. Now groups draw on both aspects: they gather not only photos on one topic but also people, who contribute (or not) to give a social identity to the group by their activity.

An Analytical Scheme

In order to sketch a map of the groups following the two aspects just described, namely tags and contacts as respectively thematic and social indicators (of course these criteria are used only as a proxy), let us present briefly two measures of these.

Given a group g, we will call the *thematic graph* (resp. *social graph*) of g the graph whose *vertices* (*i.e.* nodes) are the members of g having posted at least one photo with at least one tag, and where an (undirected) *edge* (*i.e.* link) between users u and v denotes the fact that they have at least one tag in common (resp. one is a contact of the other). Thematic edges will be weighted using a function w defined as follows.

Given a tag t and a user u, n_t and $n_t(u)$ denote respectively the number of all Flickr photos and the number of photos of user u, both having tag t (including photos outside studied groups). The maximal value of n_t is denoted by n_{max}.

The *rarity coefficient* ρ_t of a tag t is defined by $log(1+n_{max}/n_t)$. This coefficient ranges from 1 for the most used tag *beach* to approximately 10 for the rarest ones.

The *tag weight* $w_{u,t}$ of tag t on user u is defined by 0 if $n_t(u)=0$, by $1+log\ n_t(u)$ otherwise. The idea of the *log* is of course to reduce the impact of users posting thousands of photos about the same topic (their wedding, baby, cat, holiday...)

Finally the *edge weight* between users u and v is: $w_{u,v} = w_{v,u} = \Sigma_t (\rho_t \times min(w_{u,t}, w_{v,t})$, which is meant to tell whether u and v share many tags, taking into account the rarity of these tags: the rarer are the tags, the closer the users are to each other.

Most thematic groups

soc	thm	Group name
0.06	0.22	Buenos Aires
0.07	0.23	Tel Aviv Stories
0.01	0.23	K750i
0.06	0.24	Rio de Janeiro
0.05	0.24	Taipei
0.01	0.24	ROCKCLIMBING
0.08	0.24	toycamera.com
0.04	0.24	Hasselblad
0.02	0.24	Dublin
0.06	0.25	Philippines Images
0.01	0.25	Stockholm
0.03	0.25	Lisboa
0.05	0.25	XPRO CROSS PR
0.01	0.25	Vienna
0.05	0.25	Noir & Blanc
0.01	0.25	Pugs
0.03	0.25	I Shoot Fuji
0.01	0.25	Incredible India
0.03	0.25	expired film
0.03	0.25	Hamburg
0.02	0.25	Manchester, UK
0.01	0.25	snowboard
0.01	0.25	Panoramic
0.03	0.25	Baltimore
0.15	0.26	FLICKRGAYS

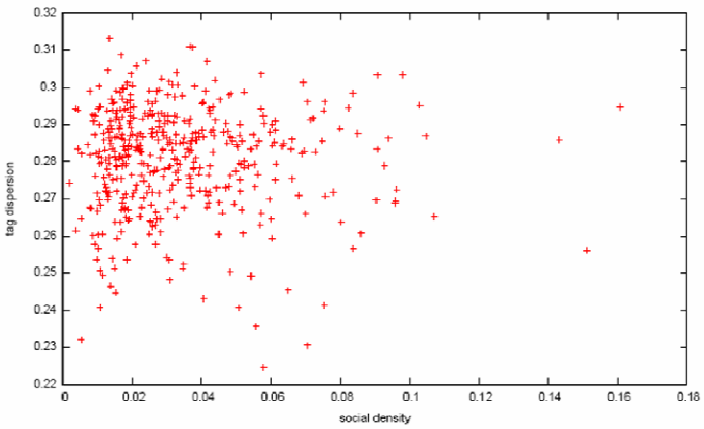

Figure 1. Social and thematic indicators for 450 groups
soc is for social density, **thm** for tag dispersion

Most social groups

soc	thm	Group name
0.16	0.29	Paralelas/Parallels
0.15	0.26	FLICKRGAYS
0.14	0.29	Fifty Faves
0.11	0.27	Pasted Paper
0.10	0.29	Nice Package!
0.10	0.30	Ethnic
0.10	0.30	poles / postes
0.10	0.27	Only 1000
0.10	0.27	handbags
0.10	0.27	mens fashion &
0.09	0.29	Let thr be light!
0.09	0.28	Desaturado
0.09	0.30	Super Colored
0.09	0.28	blue/black/white
0.09	0.27	6000+ Views
0.09	0.26	Zona libre
0.09	0.29	Male Butt Crack
0.08	0.26	Iconia Fashion
0.08	0.30	"*Alba luminis
0.08	0.29	4Variations
0.08	0.26	Handmade
0.08	0.29	Painterly
0.08	0.27	Toy Face
0.08	0.27	Photoheart
0.08	0.30	Verde - Green - Vert
0.08	0.24	toycamera.com

Let us now recall that a *Lorentz curve* graphically shows a cumulative distribution function (Figure 1 shows a Lorentz curve of all Flickr public photos, where the first 10% of the users own 70% of the photos) and that the *Gini coefficient* of a distribution is the area between the Lorentz curve and the diagonal (which is the Lorentz curve of the uniform distribution). This coefficient is a measure of the heterogeneity of the distribution: on the example, the highest numbers of photos owned by individuals are very high in comparison to photos owned by average people, the curve is thus far from the diagonal, the Gini coefficient is thus high.

We will now label a group by its *social density*, defined as the *density* of its social graph (*i.e.* the ratio of existing edges among all possible edges given the number of vertices) and its *tag dispersion*, defined as the Gini coefficient of the distribution of edge weights in its thematic graph.

Results

Figure 2 shows the results for a sample of the 450 groups having between 433 and 500 members (at the time of the crawl).

What is interesting is to look at the groups lying away from the upper-left cloud of mainstream groups with low social density and high tag dispersion. The most thematic ones, whose position is in the lower part of the chart, are listed on the left-hand side of the chart. Three-quarters of these group are in two categories: geographical, especially cities (Buenos Aires, Tel Aviv, Taipei etc.) and technical groups (K750i, XPRO, Fuji etc.), whose social densities range from very low values (Vienna, Stockholm for cities, K750i, expired films for technical) to quite high ones (Tel Aviv, Buenos Aires and toycamera, XPRO). In the case of cities, the social density may distinguish between tourism groups (where people just post photos of their travels without having much contact with others) and everyday-life groups, as suggested by the name of the group Tel Aviv Stories.

Now groups with high social density are listed on the right-hand side of the chart. Let us discuss on the first three easily distinguishable on the far right on the chart. The group Paralelas/Parallels is intended for photos with… parallel lines (wires, skyscrapers etc.), which could mean any kind of photos (the tag dispersion is high). But as suggested by the title in Portuguese, many members are from Brazil. This is an example of a social group whose social activity comes from a geographical proximity of its members (as was the case for Tel Aviv Stories). The group FLICKRGAYS is one of the (quite few) examples of both thematic and social groups and may have some relevance in terms of social cohesion. Finally, Fifty Faves is for photos having been marked as favorites by at least fifty users. Of course not thematic, this group is for very experienced Flickr users, who know each other and have discussions about their productions. In short, there is a wide range of these "social" groups, whose names and declared purposes don't necessarily tell they are social.

Conclusion

Besides showing the great diversity of uses of Flickr groups, these empirical results suggest that the methodological scheme presented in this paper may indeed be used in order to detect groups having a presumably strong social and/or thematic "identity". This could serve many purposes like targeting specific communities for designing of new services, studying how to make thematic groups become social etc.

References

Aguiton, C., and Cardon D., The Strength of Weak Cooperation: an Attempt to Understand the Meaning of Web 2.0, *Communication & Strategies*, 65, 2007.

Prieur, C., Cardon, D, Beuscart, J.-S., Pissard, N., Pons, P., The Strength of Weak Cooperation: A Case Study on Flickr, http://hal.archives-ouvertes.fr/hal-00256649/en