On the High Density of Leadership Nuclei in Endorsement Social Networks

Guillermo Garrido*
NLP & IR Group
UNED, Madrid, Spain

Francesco Bonchi
Yahoo! Research
Barcelona, Spain

Aristides Gionis
Yahoo! Research
Barcelona, Spain

ABSTRACT

In this paper we study the community structure of *endorsement networks*, i.e., social networks in which a directed edge $u \to v$ is asserting an action of support from user u to user v. Examples include scenarios in which a user u is *favoring* a photo, *liking* a post, or *following* the microblog of user v.

Starting from the hypothesis that the footprint of a community in an endorsement network is a bipartite directed clique from a set of followers to a set of leaders, we apply frequent itemset mining techniques to discover such bicliques. Our analysis of real networks discovers that an interesting phenomenon is taking place: the leaders of a community are endorsing each other forming a *very dense nucleus*.

Categories and Subject Descriptors: H.4.3 [Information Systems Applications]: Communications Applications

General Terms: Experimentation.

Keywords: Endorsement Social Networks, Communities.

1. INTRODUCTION

Understanding the viral spread of information in social media, modeling how information propagation relates to the underlying community structure, and identifying influential users, are all related tasks and important challenges with potential high returns. As a step in the direction of understanding information propagation and identifying influential users, in this paper we study the community structure of *endorsement networks*, i.e., networks in which a directed edge $u \to v$ is asserting a unit of support from user u to user v.

For instance in Flickr, a user u may *comment* or *favor* a photo of another user v. It might also be the case that u admires v's photos and wants to be updated on v's future posts: in this case u may add v as a *contact*. Indeed in Flickr contacts are unilateral, not necessarily symmetric, and they represent endorsement, not friendship. On the other hand, when a user u declares another user v as *friend* or *family*, the reason is that u wants to share her photos with v, and therefore this link represents social affinity rather than endorsement. As another example, in microblogging services such as Twitter, users post short messages which are displayed on their profile page and delivered to the author's subscribers who are known as *followers*. Being a follower is an explicit form of endorsement. In some cases a user u might "retweet" a post of user v, thus propagating the content created by v.

Analyzing endorsement networks and understanding their community structure, can lead to deeper insights in the leaders-followers relationship, and ultimately, to mastering how information and user-generated content is propagating. The applications are various, ranging from marketing and surveying, to politics and campaigning.

We start from the hypothesis that the footprint of a community in a social endorsement network is a *biclique* from a set of followers to a set of leaders. Recall that, for a bipartite subgraph formed by node sets A and B to be a biclique, every possible link from nodes in A to nodes in B must be present. Trough our analysis of real-world endorsement networks we achieve two important insights.

Large cores: endorsement networks contain large bicliques from a set of followers to a set of leaders.

Very dense nuclei: the set of leaders (nucleus) of a core almost always exhibits an extremely high internal density.

2. CORES, NUCLEI AND THEIR DENSITY

We denote the endorsement network by $G = (V, E)$, where V is a set of nodes and E is a set of *directed* edges. A directed edge $(u, v) \in E$ indicates an action of endorsement from node u to node v. A *core* $C = (L, F)$ of the network G consists of two disjoint subsets of V, i.e., $L, F \subseteq V$ with $L \cap F = \emptyset$, so that for each $u \in F$ and $v \in L$ it is $(u, v) \in E$. The set L represents the *leaders* of the core, and set F represents the *followers* of the core. The set of leaders L is also called the *nucleus* of the core. Given a core $C = (L, F)$, we define the *size* of the core $s(C)$ to be the size of the leader set L, i.e., $s(C) = |L|$, and the *support* of the core $\sigma(C)$ to be the size of the follower set F, i.e., $\sigma(C) = |F|$.

Given an endorsement network G, a threshold value s_0 on core size, and a threshold value σ_0 on core support, we seek to find all cores C in G that have size $s(C) \geq s_0$ and support $\sigma(C) \geq \sigma_0$. It is almost immediate that this is an instance of *frequent-itemset mining* [1]. Among the various strategies to deal with the patterns explosion problem, an interesting one is to consider only *maximal frequent itemsets* [2]. A maximal frequent itemset is simply an itemset which is frequent and has no frequent superset. In our context this means that given σ_0 we are not interested in a core where the nucleus of leaders is X, if the nucleus $X \cup \{v\}$ has still enough followers. The benefit of extracting only the maximal nuclei is twofold: (i) fewer and more interesting cores, and (ii) more efficient computation.

Given a core $C = (L, F)$, we define the *leader-leader density* of the core $\delta_{\mathrm{LL}}(C)$ to be the internal density of the leader set L, that is the fraction of the number of all edges between

*The author is supported by the Spanish Ministry of Science and Innovation, project QEAVis-Catiex (TIN2007-67581-C02-01).

WWW 2010, April 26–30, 2010, Raleigh, North Carolina, USA.
ACM 978-1-60558-799-8/10/04.

Table 1: Network Statistics. n: number of nodes; m: number of edges; $\bar{d}$: average degree; $\max d_{\text{in}}$: maximum in-degree; $\max d_{\text{out}}$: maximum out-degree; R: reciprocity; α_{in}: exponent of the power-law of the in-degree distribution; α_{out}: exponent of the power-law of the out-degree distribution; $\max \text{CC}$: size of the largest (strongly) connected component; $|\text{CC}|$: number of the (strongly) connected components; c: clustering coefficient.

Network	n	m	$\bar{d}$	$\max d_{\text{in}}$	$\max d_{\text{out}}$	R	α_{in}	α_{out}	$\max \text{CC}$	$\|\text{CC}\|$	c
FLICKR-E	826 829	65 851 110	79.6	22 214	15 090	0.21	1.6	1.7	486 210 (58.80%)	341 604	0.08
JAIKU	31 534	231 006	7.3	2 324	48	0.44	1.7	1.1	21 937 (69.57%)	17	0.06
FLICKR-S	687 091	10 122 046	14.7	7 610	2 867	0.48	2.1	1.8	479 127 (69.73%)	334 933	0.04
Y!360	1 921 351	7 230 996	3.8	260	260	1.00	2.5	2.5	1 463 264 (76.16%)	150 773	0.03

Table 2: For various values of s_0 and σ_0: numbers of cores found (column 3); total number of nodes which are follower (respectively, leader) in at least one core, i.e., $\mathbb{F} = \{v \mid \exists C = (F, L) \wedge v \in F\}$, and $\mathbb{L} = \{v \mid \exists C = (F, L) \wedge v \in L\}$; number of nodes that are leader in one core and follower in another one; average leader-leader and follower-follower density.

FLICKR-E

s_0	σ_0	# cores	$\|\mathbb{F}\|$	$\|\mathbb{L}\|$	$\|\mathbb{F} \cap \mathbb{L}\|$	avg δ_{FF}	avg δ_{LL}
4	90	1 267 518	22 938	2 012	1 727	0.49	0.8
4	120	65 868	10 806	653	551	0.41	0.8
4	150	5 777	4 974	198	174	0.37	0.82
5	80	3 963 545	13 079	1 407	1 176	0.60	0.89
5	90	928 484	9 631	876	731	0.54	0.87
5	100	264 548	7 303	585	485	0.51	0.87
6	80	3 203 566	6 601	740	616	0.63	0.93
6	90	630 476	4 614	442	362	0.59	0.92
6	100	145 298	3 106	241	222	0.56	0.92
6	120	7 002	1 618	92	81	0.52	0.94

FLICKR-S

s_0	σ_0	# cores	$\|\mathbb{F}\|$	$\|\mathbb{L}\|$	$\|\mathbb{F} \cap \mathbb{L}\|$	avg δ_{FF}	avg δ_{LL}
4	90	836 479	7 443	930	668	0.46	0.48
4	120	29 492	4 431	351	243	0.43	0.60
5	90	247 021	3 474	426	288	0.52	0.69
5	100	69 545	2 506	269	170	0.50	0.76
6	80	456 110	2 118	311	192	0.57	0.80
6	120	1 583	512	35	33	0.48	0.9

JAIKU

s_0	σ_0	# cores	$\|\mathbb{F}\|$	$\|\mathbb{L}\|$	$\|\mathbb{F} \cap \mathbb{L}\|$	avg δ_{FF}	avg δ_{LL}
5	50	230	135	31	12	0.49	0.93
5	30	11 218	163	80	52	0.59	0.87
4	50	250	137	32	12	0.50	0.93
4	30	13 667	848	164	115	0.59	0.86
3	50	310	993	81	37	0.44	0.86
3	30	15 132	2 260	310	227	0.57	0.84

Y!360

s_0	σ_0	# cores	$\|\mathbb{F}\|$	$\|\mathbb{L}\|$	$\|\mathbb{F} \cap \mathbb{L}\|$	avg δ_{FF}	avg δ_{LL}
4	50	8	109	8	4	0.29	0.62
4	40	66	262	25	11	0.33	0.7
5	40	1	43	5	0	0.31	0.5

nodes in L over the number of all possible edges in L:

$$\delta_{\text{LL}}(C) = \frac{|\{(u, v) \in E \mid u \in L \wedge v \in L\}|}{|L|(|L| - 1)}.$$

Similarly we define the *follower-follower density* $\delta_{\text{FF}}(C)$ to be the internal density of the follower set F.

3. EMPIRICAL FINDINGS

We analyze four datasets, two endorsement networks and two social (i.e., not endorsement) networks. The Flickr endorsement network (FLICKR-E) is a subset of the entire Flickr social network: we have a directed edge between two users u and v if user u has marked at least one photo of user v as *favorite* or if s/he has made at least one *comment* in a photo of v. Our second endorsement network is Jaiku (JAIKU), a *micro-blogging* social network. Here we have a directed edge from user u to user v whenever user u is *following* user v. The Flickr social network (FLICKR-S), use the same sample of users as in the case of FLICKR-E, but in this case a directed edge from users u to user v indicates that user u has marked user v to be their *"friend"* or *"family"*. The second social network we use is Yahoo! 360 (Y!360), an undirected network that indicates friendship relationship among users. This is the unique undirected network we use, but we can make it directed by considering for each edge the two links in both directions. The basic characteristics and statistics of our datasets are reported in Table 1. Notice that the JAIKU network is significantly smaller than the other three, on the other hand, the Y!360 network is the sparsest of all. Note that although the networks FLICKR-E and FLICKR-S are defined over the same base of users, they have different number of nodes due to the removal of singleton nodes.

We next report the empirical evidence of our findings, namely that large cores can be found in endorsement networks and that these cores have a very dense leadership nucleus. Indeed, our results (reported in Table 2) clearly show that δ_{LL} is usually very large for endorsement networks, while it is always smaller for friendship-based social networks. In both endorsement and social networks, the average density of links among the followers (i.e., δ_{FF}) is always much lower than the nucleus density (i.e., δ_{LL}). This clearly shows the presence of a strong *directionality* of the links: mainly from the followers to the leaders. Recall that $\delta_{\text{FL}}(C) = 1$ by definition, or in other terms, in a core all followers point to all leaders. It is worth mentioning that we can not use the same settings of the parameters s_0 and σ_0 in all the networks, as they have different sizes and different densities: what is a reasonable settings for one network could result in too few cores in another network.

Using the method of *swap randomization* we confirm that the structure of the cores that we report in this paper is statistical significant.

Finally, as it is usually the case when mining any form of frequent patterns, our method produces many similar, overlapping, redundant cores, which presumably are different footprints of the same community. This indicates the need to devise clustering technique in order to coalesce similar cores into meaningful communities, having a very large followers base, while still maintaining a very high density in their leadership nucleus.

4. REFERENCES

[1] R. Agrawal, T. Imielinski, and A. N. Swami. Mining association rules between sets of items in large databases. In *Proceedings of ACM SIGMOD* 1993.

[2] R. Bayardo. Efficiently mining long patterns from databases. In *Proceedings of ACM SIGMOD* 1998.