What Women Like: A Gendered Analysis of Twitter Users' Interests based on a Twixonomy

Stefano Faralli, Giovanni Stilo and Paola Velardi
Department of Computer Science
Sapienza University of Rome, Italy
lastname@di.uniroma1.it

Abstract

In this paper we analyze the distribution of interests in a large population of Twitter users (the full set of 40 million users in 2009 and a sample of about 100 thousand New York users in 2014), as a function of gender. To model interests, we associate "topical" friends in users' friendship lists (friends representing an interest rather than a social relation between peers) with Wikipedia categories. A word-sense disambiguation algorithm is used for selecting the appropriate wikipage for each topical friend. Starting from the set of wikipages representing the population's interests, we extract the sub-graph of Wikipedia categories connected to these pages, and we then prune cycles to induce a direct acyclic graph, that we call Twixonomy. We use a novel method for reducing the computational requirements of cycle detection on very large graphs. For any category at any generalization level in the Twixonomy, it is then possible to estimate the gender distribution of Twitter users interested in that category. We analyze both the population of "celebrities", i.e. male and female Twitter users with an associated wikipage, and the population of "peers", i.e. male and female users who follow celebrities.

1. Introduction

In this paper we present a method for extensively analyzing the distribution of interests in Twitter according to gender. Our work is related with two areas in social media analytics: analysis of users' interests and gender studies. Large-scale studies of Twitter users across the world mainly report simple demographic statistics[1] like gender, age and geographic distribution, followers and following counts, etc. A considerable number of works are aimed at modeling users' interests for some specific purpose, like detecting trending topics, i.e. topics that emerge and become popular in a specific time slot. Trending topics are extracted to model users' expertise (Wagner et al., 2012), to produce a recommendation (Garcia and Amatriain, 2010; Kywe, Lim, and Zhu, 2012; Lu, Lam, and Zhang, 2009)[2], or to analyze general interests (e.g. events) that are predominant in a given time span (Li et al., 2010). The majority of these methods infer interests from lexical information in tweets (bigrams, named entities or latent topic models), a technique that may fall short in terms of computational complexity when applied to large Twitter populations, as shown in Stilo and Velardi (2014).

Only few studies investigated the characteristics of Twitter users regardless of specific applications. In Kim et al. (2010) it is shown that words extracted from Twitter lists could represent latent characteristics of the users in the respective lists. In Kapanipathi et al. (2014) named entities are extracted from tweets, then, Wikipedia categories, named *primitive interests*, are associated to each named entity. To select a reduced number of higher-level categories, named *hierarchical interests*, spreading of activation (Anderson, 1968) is used on the Wikipedia graph, where active nodes are initially the set of primitive interests. Note that, despite their name, hierarchical interests are not hierarchically ordered. Furthermore, as discussed later in this paper (Section 5), higher level categories in Wikipedia may be totally unrelated with some of the connected wikipages.

Similarly to us, Bhattacharya et al. (2014) try to infer users interests at a large scale. Their system, named Who Likes What, is the first system that can infer users' interests in Twitter at the scale of millions of users. First, the topical expertise of popular Twitter users is learned using a latent model on Twitter lists in which such users actively participate. Then, the interests of the users following through the lists such expert users are transitively inferred. By doing so, Who Likes What can infer the interests of around 30 millions users, covering 77% of the analyzed populations. Evaluation is performed at a much smaller scale, by manually comparing extracted interests with those declared in a number of users' bio, and by using human feedback from 10 evaluators. The evaluators commented that the inferred interests, even though useful, are sometimes too general: on the other side, given the large and unstructured nature of the extracted interests (over 36 thousand distinct topics), generating labels at the right level of granularity is not straightforward.

Concerning gender studies, research mainly concentrated on gender profiling, i.e. automatically inferring a user's gender (Marquardt et al., 2014; Sap et al., 2014; Smith, 2014), and on the analysis of gendered language online (Bamman, Eisenstein, and Schnoebelen, 2014). An interesting work

[1] http://www.beevolve.com/twitter-statistics

[2] The literature on interest-based user recommendation is very vast and it would be impossible to survey it all. Refer to Kywe, Lim, and Zhu (2012) for a survey.

(Szell and Thurner, 2013) has been recently published in which the authors analyze gender differences in the social behavior of about 300,000 players of an online game. To the best of our knowledge, no studies on users' interests in social network by gender have been published so far, except for a sociological analysis on women's soccer Twitter audience (Cochea, 2014).

With respect to the analyzed bibliography, our main contribution is the acquisition of a Twixonomy, a directed acyclic graph (DAG) of Twitter users' interests, inferred from users' friendship information of the entire Twitter population, and from Wikipedia. We prefer friendship rather than textual features to model users' interests, since, unless we are addressing a specific community (like e.g. the members of a political party, as in Colleoni, Rozza, and Arvidsson (2014)), the number of inferred topics may quickly grow, as in Bhattacharya et al. (2014) , and it is very hard to make sense of them, or even to evaluate their quality. Furthermore, textual features such as word clusters are temporally unstable as compared to friendship and categorial interests, as already shown in Myers and Leskovec (2014) and Siehndel and Kawase (2012). In Barbieri, Manco, and Bonchi (2014) the authors argue that users' interests can be implicitly represented by the authoritative users (named hereafter topical users) they are linked to by means of friendship relations. This information is available in users' profiles, and does not require additional textual processing. Topical friends are therefore both stable and readily accessible indicators of a user's interest. However, as a mean to systematically analyze interests in large networks, this information is hardly interpretable and sparse, just like lexical features and lists.

To obtain a hierarchical representation of interests, we first associate a Wikipedia page with topical users in users' friendship lists. We denote as *topical users* those for which one such correspondence exists. This definition is slightly different from that adopted in Barbieri, Manco, and Bonchi (2014)[3], however it seems equally intuitive.

In general many pages can be associated with a Twitter account name, therefore we use a word sense disambiguation algorithm, as detailed later in this paper. Users can then be directly (if they map to a wikipage) or indirectly (if they follow a mapped user) linked to one or more Wikipedia pages representing his/her $primitive$ interests, (we use the same terminology as in Bhattacharya et al. (2014)). Starting from the set of wikipages representing the population's interests, we consider the Wikipedia category sub-graph G induced from these pages, and we then remove cycles to obtain a direct acyclic graph that we call Twixonomy. Efficient cycle pruning on the very large graph G is performed using an iterative algorithm.

Since every node in the Twixonomy can be associated with the set of Twitter users hierarchically linked to that node via their primitive interests, our Twixonomy is helpful for a variety of tasks, such as hierarchically-tunable user profiling, community detection in selectable domains, and sociological analysis, like the study of gender interests by category, which is the focus of this paper. In our gender analysis we consider both celebrities (i.e. male and female topical users[4]) and male and female common users.

The paper is organized as follows: In Section 2 we shortly describe the datasets and tools used in this study, Section 3 presents the algorithm to create the Twixonomy, Section 4 is dedicated to a comparison with Who Likes What (Bhattacharya et al., 2014) and Section 5 performs a study of gender distribution across categories. Finally, Section 5 is dedicated to concluding remarks and future work.

2. Data and resources

For our study we use the following resources:

- **The Twitter 2009 network** The authors in Kwak et al. (2010) have crawled and released the entire Twitter network as of July 2009. Since Twitter data are no longer available to researches, this remains the largest available snapshot of Twitter, with 41 million user profiles and 1.47 billion social relations. Even though things might have changed in Twitter since 2009 - the number of users has grown up to 500 millions - our purpose in this paper is to demonstrate the efficiency of our algorithms on a very large sample of users.
- **The Twitter 2014 NewYork network** On June 2014 we crawled a sample of New York Twitter users starting from a seed of 3800 users who tweeted more than 20 times in New York[5]. With respect to the Twitter 2009 dataset, this network is much smaller but highly connected.
- **Babelfy** Babelfy (Moro, Raganato, and Navigli, 2014) is a graph-based word-sense disambiguation (WSD) algorithm based on a loose identification of candidate meanings coupled with a densest sub-graph heuristic which selects high-coherence semantic interpretations. Babelfy disambiguates all nominal and named entity mentions occurring within a text, using the BabelNet semantic network (Navigli and Ponzetto, 2012) a very large multilingual knowledge base, obtained from the automatic integration of Wikipedia and WordNet. Babelfy has shown to obtain state-of-the-art performances in standard WSD benchmarks and challenges. Both BabelNet and Babelfy are available online[6] .
- **The Wikipedia Graph** We created the Wikipedia graph from the Wikipedia dump in 2009 and 2014 (for coherence with the two Twitter population datasets). The Wikipedia graph is the basis from which we infer the Twixonomy for each of the two Twitter populations.

3. The Twixonomy

This Section describes the algorithm to obtain the Twixonomy starting from a Twitter population P. First, we extract

[3] in this paper topical users are identified according to the structural properties of the network

[4] several Twitter accounts correspond to organizations, places, events and products rather than individuals

[5] the details of the geo-localization algorithm are omitted for the sake of space and because they are outside the scope of the paper

[6] http://babelnet.org/, http://babelfy.org/

Algorithm 1 Build Twixonomy

Input: F = twitter users followed by at least one member of the initial Twitter population P
CG: top category hierarchy from Wikipedia

Output: a DAG taxonomy where:

- leaf nodes are twitter user mapped into wikipages , and the remaining nodes are Wikipedia categories;
- edges are one of three kinds: <super-category , category>, <category , wikipage>, <wikipage , Twitter "topical" user>

```
 1: G = empty directed graph
 2: for each twitter u:F do
 3:    u.senses = ∅;
 4:    u.profile = Twitter.getProfile(u);
 5:    senses = BabelNetSenses(u.profile.name);
 6:    if |senses|==1 then
 7:       u.senses = senses
 8:    else
 9:       target = u.profile.name;
10:       context = {
             u.profile.name,
             u.profile.statusline,
             u.profile.location
          };
11:       u.senses = Babelfy.getSenses(target,context);
12:    end if
13:    for each sense ∈ u.senses do
14:       G.addEdge( sense , u.profile.screenName );
15:       for each edge ∈ path(sense,CG) do
16:          G.addEdge(edge);
17:       end for
18:    end for
19: end for
20: TWIXONOMY= removeCycles(G);
21: return TWIXONOMY;
```

from users' profiles the set F of users followed by at least one user in P. Note that the sets P and F are different, though possibly overlapping: for Twitter 2009, since P is the complete Twitter population, we have $F \subset P$ and for the NY population we have instead $|F| \gg |P|$. We generate the Twixonomy from the set F, as explained in what follows, with reference to the pseudo-code shown in Algorithm 1 and Algorithm 2.

3.1 Identify topical nodes

For every user $u \in F$, the objective is to identify a corresponding wikipage in Wikipedia, if any. As we already clarified, "topical users" are those u for which one such correspondence exists. Obtaining a correspondence between a user screen name and a Wikipedia category e.g. @britneyspears → Britney Spears, is not trivial for a number of reasons. First, Twitter names do not straightly correspond to Wikipedia page names, and secondly, many pages can be associated to a named entity, for example: Britney (person), Britney (album), Britney (Busted song), Britney ("For the Record" documentary), etc. We perform joint name resolution and disambiguation (in case of multiple corresponding nodes) using Babelfy (Moro, Raganato, and Navigli, 2014), which disambiguates a textual input against BabelNet

Algorithm	Worst case time complexity
Tiernan (1970)	$O(V^V)$
Tarjan (1972)	$O(V * E * C)$
Johnson (1975)	$O((V + E) * C)$
J.L.Szwarcfiter and P.E.Lauer (1974)	$O(V + E * C)$

Table 1: Summary of the worst case time complexity of the algorithms for finding cycles in directed graphs (C = number of cycles, V = number of nodes and E = number of edges.

senses[7]. For any user and screen-name, e.g. @britneyspears, we first retrieve from the corresponding Twitter profile the fields $name$, $line - status$, and $location$, e.g. "Britney Spears", "Its Britney ...", "Los Angeles, CA". Then, we retrieve all BabelNet senses associated to the $name$ field (lines 4-5 of Algorithm 1) and, if there are multiple senses, we submit to Babelfy the sentence generated by concatenating these strings, e.g. "Britney Spears It's Britney ... Los Angeles, CA". Finally, we retrieve the disambiguated sense(s) that Babelfy has associated to the string $name$. These steps are shown in lines 5-12 of Algorithm 1. With reference to our previous example, the sense $Britney\ (person)$ is returned. Note that in many cases there are no senses corresponding to a Twitter $name$ field, as expected, since most users in F are common users. In some cases however a match would exist but is missed, e.g @pinballwizard (i.e. pinball wizard, whose $name$ field is again the non splitted $pinballwizard$). To increase the recall, we use a name splitting heuristics when no BabelNet senses are retrieved from the $name$ field (this step is omitted in Algorithm 1 for the sake of brevity).

3.2 Build the Twixonomy

Let's denote with T the set of wikipages associated with the topical users in F: these represent the "leaf nodes" [8] of the Twixonomy. Note that, after disambiguation, there is one leaf node (i.e. a wikipage) for each topical user in T. Furthermore, every node $t \in T$ is associated with the number of users in P who follow t.

We then consider in the Wikipedia graph all the nodes that can be reached starting from any $t \in T$ and traversing the graph up to one of the 22 Wikipedia top categories[9], i.e. Art, Agriculture, Concepts, etc (these steps are shown in lines 13-17 of Algorithm 1). The resulting graph G, even starting from a relatively small population P (like the NY-Twitter 2014), is still very large (since T can be quite large), and furthermore has a high number of cycles[10], e.g. Economics lists → Business lists → Economics lists. To obtain a DAG (directed acyclic graph), i.e. our final Twixonomy, we need to

[7] remember that BabelNet senses are mapped to Wikipedia pages

[8] hereafter we define these nodes interchangeably as as topical nodes, leaf nodes, or wikipages

[9] http://en.wikipedia.org/wiki/Category:Main_topic_classifications

[10] http://en.wikipedia.org/wiki/Wikipedia:Dump_reports/Category_cycles

Algorithm 2 Remove Cycles

Input: a directed GRAPH G
Output: a DIRECTED ACYCLIC GRAPH (DAG)
1: **while** $(VC = detectCycle(G)) <> \emptyset$ **do**
2: $G' = G[VC]$ (vertex-induced subgraph of G)
3: cyc = getOneCycle(G')
4: break the cycle cyc on G;
5: **end while**
6: **return** G

	Twitter 2009	NY-Twitter 2014
#users (P)	40,171,624	101,362
#topical users (T)	1,787,909	736,929
Average ambiguity of topical users before disambiguation	5.27	5.33
% of users described by at least one topic	66%	99%

Table 2: Network statistics

remove cycles.

There are several algorithms for identifying simple cycles in graphs, like those listed in Table 1 along with their worst case complexity formulas. In practice, all these algorithms have a high administrative cost in terms of time and memory, therefore we defined an optimized iterative algorithm.

The procedure to remove cycles, based on topological sorting (Kahn, 1962), is summarized in Algorithm 2. In line 1, the $detectCycle$ procedure is iteratively applied on a graph G. Our procedure, based on Kahn's topological ordering algorithm, returns the set of nodes VC in G belonging to at least one cycle. This is obtained by ordering the nodes of the directed graph G and identifying cases for which topological ordering is not possible because there is a cycle. This step has a complexity of $O(V+E)$ (Kahn, 1962). Then (line 2) we consider the vertex-induced subgraph G' of the set VC, and we apply the $getOneCycle$ procedure. This procedure, again based on topological ordering, returns the first encountered cycle in G', which is subsequently broken in G (lines 3-4). Steps 1-4 are iterated on the reduced graph, until no more cycles are found. Overall, the worst case complexity is $O((V+E)*C)$, where C is the number of cycles in G.

Even though the worst case complexity of Algorithm 2 is the same as for Johnson's algorithm (see Table 1), an optimized use of computational resources derives from the fact that in general $G' \ll G$, and that topological ordering has reduced memory requirements with respect to "classical" cycle detection algorithms. In practice, on the very large Wikipedia graph obtained when starting from the Twitter 2009 population, the algorithm was able to remove all cycles in 12 hours, while the algorithms in Table 1 either saturated the memory or could not return a solution after six days, using a mid-high level desktop computer.

Table 2 shows some network statistics. In the Twitter 2009 dataset, we identified 1,8 million topical users and in the NY-Twitter 2014 dataset over 700 thousand topical users, even though the initial population P is two orders of magnitude smaller than for Twitter 2009. Figure 1 shows the coverage of the Twitter 2009 and NY-Twitter 2014 populations as a function of the number of expressed interests. The two populations are rather different in this respect: in the 2009 dataset 66% of the population P is described by at least one topic (and related categories), while e.g. 5% is described by at least 20 topics. Instead, 99% of NY-Twitter 2014 is described by at least one topic and 80% has at least 20 topics. NewYorkers are considerably more connected with respect to the "older" 2009 network, both because rapidly increasing connections is a general trend in the Twitter graph, and because this is a tendency of NY citizens[11].

Concerning coverage, Figure 1 favorably compares with the results in Bhattacharya et al. (2014), where the authors mention that their coverage is 77% on a network sample which also dates 2014. In their system, however, interests are $induced$ from those of expert users, rather than $explicitly$ mentioned in a user's profile, therefore in princi-

	Twitter 2009	NY-Twitter 2014
#nodes in G before pruning	3,146,851	1,542,924
#links in G before pruning	5,628,750	3,397,353
#nodes in pruned Twixonomy	2,195,441	1,038,205
#links in pruned Twixonomy	3,202,959	1,863,286
Max depth of Twixonomy	15	15

Table 3: Twixonomy statistics

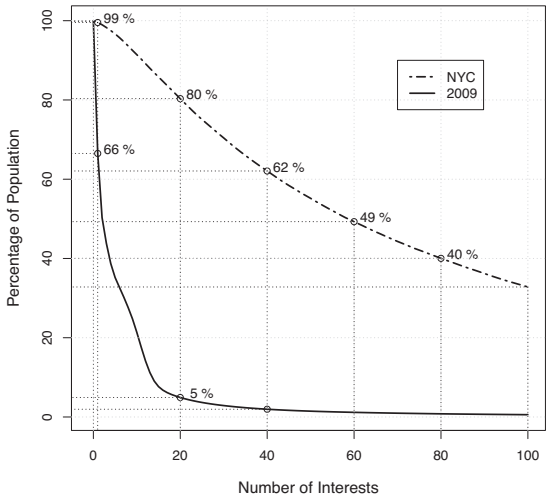

Figure 1: Coverage as a function of the number of detected topics per user (Twitter 2009 and NY-Twitter 2014)

[11] http://www.statista.com/statistics/322947/facebook-fans-twitter-followers-of-new -york-knicks/

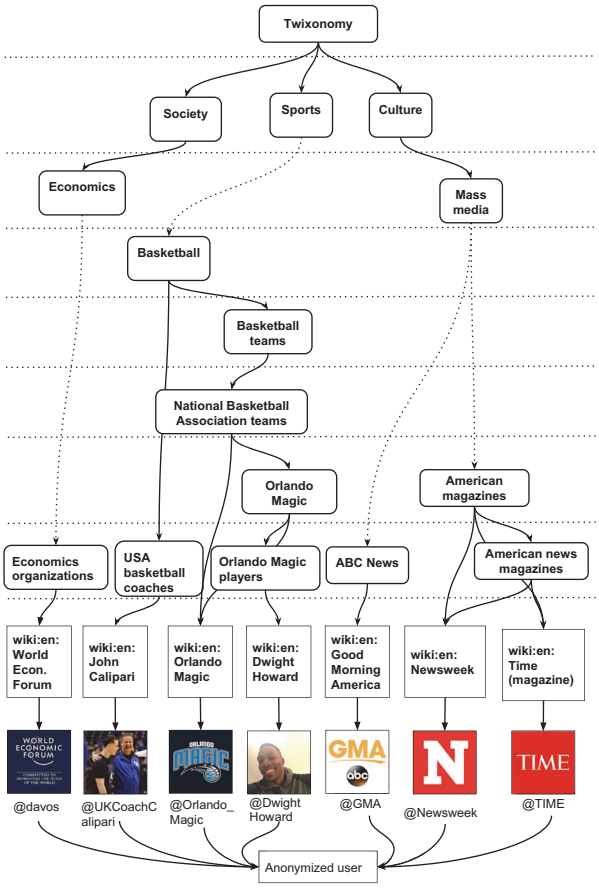

Figure 2: Example of Twixonomy for a single user

ple our methodology is also more reliable. We note that, to further improve coverage, we could use a method similar to Bhattacharya et al. (2014), inferring additional interests for a user, based on his/her peer friends. A study on interests propagation in communities is left to future work.

Table 2 also shows that the initial ambiguity of topical users' names was rather high (5.27 for Twitter 2009 and 5.33 for NY-Twitter 2014). Though Babelfy has been extensively evaluated in Moro, Raganato, and Navigli (2014), we manually evaluated a sample of 200 ambiguous user names for which a wikipage was selected by Babelfy, and 200 names for which no correspondence was found, achieving an F-measure of 0.82. To improve precision, similarly with what we proposed for coverage, topical users' peer friends profiles could be used to provide Babelfy with more context.

Table 3 shows some Twixonomy statistics, such as the number of nodes and edges before and after removing cycles, and the max depth of the extracted Twixonomy. We can see that, even starting from very different population sizes, the two Taxonomies are of the same order of magnitude.

Note that with the same method illustrated in Algorithm 1 we can build a single-user or a community Twixonomy. For example, Figure 2 shows the Twixonomy of a single "common" user with 7 topical nodes in his/her friendship list. The Figure shows (along with other examples that we analyzed) that mid-general categories are the most representative of a user's interests since, as the distance between a wikipage and a hypernym node increases, the semantic relatedness decreases. In the example, the categories *Economics, Basketball* and *Mass Media* could be chosen to cover of all user's topical friends.

Our Twixonomy is made available on http://anonymous-submission-required, along with the set of users' IDs in P connected with the Taxonomy nodes.

4. Comparison with Who Likes What

So far we already compared our system with Who Likes What (Bhattacharya et al., 2014), highlighting two advantages of our Twixonomy:

- a hierarchical organization of interests, rather than an unstructured and large set of topic labels;
- a higher coverage, achieved by extracting explicit users' interests rather than induced interests.

In this section we perform a more detailed analysis of the differences between the two systems. Who Likes What (WLW) is accessible from http://twitter-app.mpi-sws.org/who-likes-what/. In this web site, it is possible to visualize the interests of a users (in the form of an ordered list of topics or a topic cloud) by providing his/her name. It is also possible to inspect a number of sample interests of Media Personalities, Researchers and Geeks.

Figure 3 is the tag cloud of the first personality shown in the Media list, Nathan Fillion, precisely as shown in WLW. We created the Fillion's Twixonomy from his set of topical friends, and we then generated a set of category *k-lists*, such that each list includes the k-hop level categories (the categories that are reached in k hops from Fillion's leaf topical nodes), weighted by the out-degree of the node. Figures 4 and 5 show the tag clouds of the 1-list and 2-list , respectively.

There is no easy way to quantitatively compare the performances of the two methods "in the large", given the different vocabulary (plain English words against Wikipedia categories) however comparing the tag clouds it is seen that categories, especially mid-low level ones, represent a more intuitive and precise description of Fillion's interests than WLW topics. In particular, the first two categories in Figures 4 (American Film Actors and American Television) summarize the majority of WLW topics in Figure 3, i.e.: *celebrities, celebs, entertainment, movies, actors, famous, tv, actresses, film, stars, hollywood, television, comedians, artists*, but are quite more specific. As the level of generality increases, e.g. in Figure 5 where k=2, the similarity between the WLW and Twixonomy clouds increases in terms of tag matches, however, as also remarked by the WLW evaluators, the interests, even though useful, become too general. Similar results have been obtained for all 17 sample interests accessible in the WLW web site.

5. What Women like?

The main advantage of the Twixonomy is that we are able to describe the interests of single users, communities, or the

Figure 3: WLW interest cloud of @NathanFillion http://twitter-app.mpi-sws.org/who-likes-what/sample-interests.php?group=media_users

Figure 4: Interest cloud derived from 1-hop categories in the @NathanFillion Twixonomy.

Figure 5: Interest cloud derived from 2-hop categories in the @NathanFillion Twixonomy.

entire network at selected levels of granularity, as in the examples of Figures 4 and 5. In this paper, our aim is to perform a gender analysis of Twitter users' interests, but many other applications are possible. To identify gender, we used a large list of female and male names extracted from several available sources[12] . More complex algorithms can be used,

[12]e.g.: http://en.wikipedia.org/wiki/Category:Given_names; http://babynames.net/ etc.

	Twitter 2009		NY-Twitter 2014	
	Topical users (T)	Users (P)	Topical users (T)	Users (P)
males (M)	829,565	13,554,883	357,002	39,871
females (F)	312,190	10,849,637	156,603	29,264
not gendered (U)	646,152	15,767,104	223,304	32,227
F/(F+M)	27.3	47.6	30.5	44.6
Average female interests	n/a	8.9	n/a	95.5
Average male interests	n/a	9.4	n/a	103.9
Female interests(%)	n/a	44.2	n/a	40.7
Male interests(%)	n/a	55.8	n/a	61.3

Table 4: Gender statistics

like gendered language and other gender signals in a user's profile, however on a very large population this is computationally demanding. We computed precision and recall of gender classification on a sample of 200 names, and we obtained an F-measure of 85%. Importantly, errors are independently distributed and do not significantly alter the gender statistics. We are aware, however, that women more than men are concerned with privacy and declare their accounts as private [13].

In our study, we aim to analyze two distinct populations: common users and topical users. As we already remarked, several Twitter accounts do not correspond to individuals, but rather, to organizations, products, places, etc. Furthermore, a number of names cannot be reliably associated with a gender. Therefore "gendered" users are a subset of both topical and common users. Figure 6 clarifies the different types of populations we are dealing with: P is the initial set of users, who can be male (M) female (F) or other (U). The set of topical users T is also partitioned in the same three categories. Hereafter we refer to gendered topical users as to $celebrities$, and to gendered common users, as to $peers$ to avoid confusion with the respective full network populations T and P. Furthermore, since users in P may express several interests, or none, as shown in Figure 6, the number of, e.g. *peer women's interests* FT feeding the Twixonomy is different from the number of *peer women* F interested in at least one category of the Twixonomy. Similarly, the number of *peer men's interests* MT is different from the number of *peer men* M.

Table 4 shows the gender distribution of celebrities, peers, and peer's interests, for both the Twitter 2009 and NY-Twitter 2014 datasets. It is immediately seen in Table 4

[13]http://www.beevolve.com/twitter-statistics/#b1

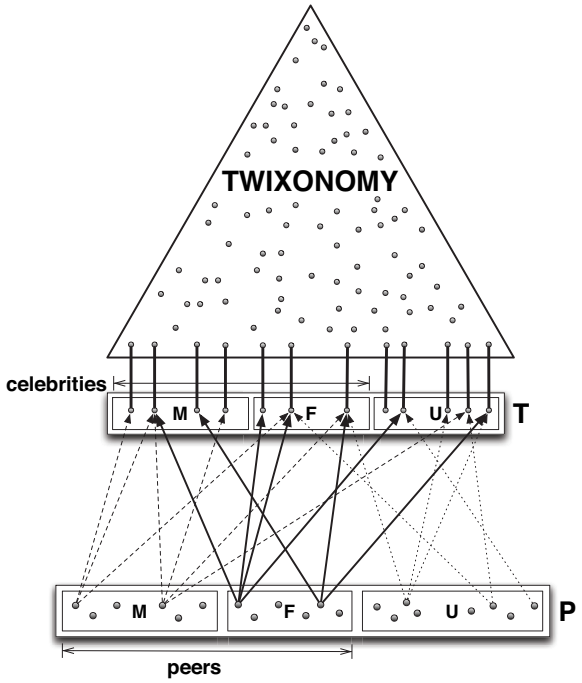

Figure 6: Mapping scheme between Twitter users and the Twixonomy

that the percentage of female celebrities is considerably less than female peers, and both are less than males. Furthermore, even though female peers are around 44-47% in the two datasets, since women tend to express in the average a lower number of interests, the percentage of female interests is slightly lower than the percentage of female peers. Comparing the two datasets, the main difference is in the average number of interests expressed by peer users, which is one order of magnitude higher for NY-Twitter, a difference that we already motivated in Section 3.

We first analyzed the distribution of celebrities and peers' interests in the Twixonomy topmost categories. The statistical significance of all the results reported hereafter has been tested using the chi-square test (Cochran and Snedecor, 1989) and the web application in http://graphpad.com/quickcalcs/chisquared1.cfm. We found that the proportion of celebrities and peers' interests in topmost categories is not statistically significant as compared with the respective proportion in full populations, except for the category Sports, where males dominate. This is also due to the fact that there are too many paths in Wikipedia (and in the Twixonomy) trough which topmost categories can be reached, some of which are rather unexpected, e.g. *Mathematics* $\rightarrow$ *Theoretical computer science* $\rightarrow$ *Statistics* $\rightarrow$ *Kindship and descent* $\rightarrow$ *Genealogy* $\rightarrow$ *Given names*, a path connecting any given name with $Mathematics$. This example, and many similar ones, confirms that mid-low categories are a better description of users' interests since, after a number of hops from the initial wikipage, the reached categories can be totally unrelated. More interesting results are instead obtained with lower level categories, as already discussed with reference to Figures 2, 4 and 5.

Table 5 shows the results for some[14] of the mid-general categories for which we found a statistically significant difference (either for celebrities or peers' interests) with respect to the full populations. We observe that, in both datasets, female celebrities are Women Organizations' and Fashion's leaders and, in Twitter 2009, also Pop Musicians. We also note that there are more Democrats than Republicans and more Democrats' followers than Republicans' followers (a difference that has been observed also in the full Twitter population (Colleoni, Rozza, and Arvidsson, 2014)).

In general, there is an agreement between the percentage of celebrities and celebrities' followers, in the sense that categories in which the percentage of female celebrities is higher (with reference to the average value) are also categories in which the percentage of female interests are higher, with some exception: for example, though only 4-5% Current National Leaders are female (well below the average of female celebrities in the Twixonomy), the percentage of female's interests in this category is more or less in the average (i.e. not significantly diverging from the fraction of female interests in the full population). In other terms, there are very few women leaders, but women are indeed enough interested in leadership: it seems however that they prefer to follow male leaders, as shown in Table 6, in which we measure the degree of homophily for each category c of Table 5.

Homophyly is computed as the ratio between the number of female interests in female celebrities FFT_c and the total number of female interests FT_c in the topics of the category. Note that FT_c includes interests in female celebrities, male celebrities and also "other" non-gendered topics UFT_c, therefore we have that e.g.: $(FFT_c + MFT_c + UFT_c/FT_c) = 1$. The Table shows that men have a significantly higher tendency towards homophily than women. Note that significance in a category must be tested against the distribution of female and male celebrities in that category: for example, if there are 5.4% female celebrities in Current national leaders, the expected fraction of peer female interests in female leaders should be close to that value, in absence of homophyly. Instead, we note that except for the categories Writers, Democrats and Women's organizations, women are either non-homophylous or they support man or non-gendered entities significantly more than other women.

Overall, the results obtained for the two datasets, in spite of the temporal distance, are remarkably in agreement, except, of course, for the absolute numbers. This is bad news, since there are no perceivable changes in the degree of predominance of males, especially as far as celebrities and traditional male's domains are concerned.

6. Concluding remarks and future work

In this paper we described a novel method to induce a Twixonomy (a hierarchical representation of Twitter

[14] for the sake of space we can only present an excerpt of our results, however, as previously mentioned, the Twixonomy is available along with the set of peer users' IDs in each category.

	%Female Celebrities		Peers %Female interests	
	T-09	N-T-14	T-09	N-T-14
Avg. Population values	27.3	30.5	47.6	44.6
Pop musicians	61.6	23.0	58.3	47.6
Schoolteachers	33.3	31.8	56.9	37.4
Writers	23.8	25.0	45.7	38.8
Businesspeople	22.4	28.9	47.3	40.3
Sportspeople	10.1	10.6	39.5	29.0
Current national leaders	4.2	5.4	48.3	41.9
Religious leaders	9.3	11.4	40.6	37.8
Fashion	56.9	63.1	54.7	50.6
Women's organizations	79.7	66.9	58.8	49.7
Military organization	12.1	13.4	39.9	32.8
Democrats (United States)	18.6	20.0	49.2	40.6
Republicans (United States)	11.1	10.1	44.5	35.1
Category for which the % of females is higher	Women's organizations	Women's organizations	Women's organizations	Fashion
Category for which the % of females is lower	Current national leaders	Current national leaders	Sportspeople	Sportspeople

Table 5: Mid-general categories in Twixonomy for which there is a statistically significant difference in the distribution of celebrities and peers in Twitter 2009 (T-09) and NY-Twitter 2014 (N-T-14).

	T-2009		T-NY-2014	
	MMT/MT	FFT/FT	MMT/MT	FFT/FT
Pop musicians	37.9	63.1	63.6	36.6
Schoolteachers	8.2	21.0	62.2	43.5
Writers	76.0	32.0	76.0	33.7
Businesspeople	83.2	27.0	79.9	30.3
Sportspeople	85.6	24.5	89.5	16.4
Current national leaders	99.1	0.4	98.5	1.0
Religious leaders	89.0	17.8	82.6	21.2
Fashion	35.3	61.4	42.8	58.6
Women's organizations	0.2	99.8	42.7	58.8
Military organization	96.8	2.8	92.7	8.3
Democrats (United States)	67.1	45.5	76.5	29.4
Republicans (United States)	95.9	7.2	85.4	22.2

Table 6: Homophyly degree in categories

users'interests), based on Wikipedia categories. A Twixonomy can be induced for single users, communities, and populations, thus providing material for a variety of demographic analyses. We applied the Twixonomy to the study of gendered interests in two large Twitter populations, that led to a number of interesting findings.

Our work can be extended in many ways: the quality and coverage of the Twixonomy can be further improved by exploiting the network structure both to increase precision of Twitter names sense disambiguation and coverage of users; a more systematic analysis of the best generalization level to describe users' interests can be conducted; pruning strategies to delete less meaningful Wikipedia hypernymy relations in the Twixonomy can be devised, and more.

References

Anderson, H. 1968. Fire spread and flame shape. *Fire Technology* 4(1):51–58.

Bamman, D.; Eisenstein, J.; and Schnoebelen, T. 2014. Gender identity and lexical variation in social media. *Journal of Sociolinguistics* 18(2):135–160.

Barbieri, N.; Manco, G.; and Bonchi, F. 2014. Who to follow and why: Link prediction with explanations. In *Proceedings of The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2014)*.

Bhattacharya, P.; Zafar, M. B.; Ganguly, N.; Ghosh, S.; and Gummadi, K. P. 2014. Inferring user interests in the twitter social network. In *Proceedings of the 8th ACM Conference on Recommender Systems*, RecSys '14, 357–360. New York, NY, USA: ACM.

Cochea, R. 2014. *Soccer and Society* 15:449–471.

Cochran, W. G., and Snedecor, G. W. 1989. Iowa State University Press.

Colleoni, E.; Rozza, A.; and Arvidsson, A. 2014. Echo chamber or public sphere? predicting political orientation and measuring political homophily in twitter using big data. *Journal of Communication, Special Issue: BIG DATA IN COMMUNICATION RESEARCH* 64:2.

Garcia, R., and Amatriain, X. 2010. Weighted content based methods for recommending connections in online social networks. In *The 2nd ACM Workshop on Recommendation Systems and the Social Web*.

J.L.Szwarcfiter, and P.E.Lauer. 1974. Finding the elementary cycles of a directed graph in o(n + m) per cycle. 60.

Johnson, D. B. 1975. Finding all the elementary circuits of a directed graph. *SIAM J. Comput.* 4(1):77–84.

Kahn, A. B. 1962. Topological sorting of large networks. *Commun. ACM* 5(11):558–562.

Kapanipathi, P.; Jain, P.; Venkataramani, C.; and Sheth, A. 2014. User interests identification on twitter using a hierarchical knowledge base. In Presutti, V.; d'Amato, C.; Gandon, F.; d'Aquin, M.; Staab, S.; and Tordai, A., eds., *The Semantic Web: Trends and Challenges*, volume 8465 of *Lecture Notes in Computer Science*. Springer International Publishing. 99–113.

Kim, D.; Jo, Y.; Moon, I.-C.; and Oh, A. 2010. Analysis of twitter lists as a potential source for discovering latent characteristics of users. In *CHI 2010 Workshop on Microblogging*.

Kwak, H.; Lee, C.; Park, H.; and Moon, S. 2010. What is Twitter, a social network or a news media? In *WWW '10: Proceedings of the 19th international conference on World wide web*, 591–600. New York, NY, USA: ACM.

Kywe, S.; Lim, E.-P.; and Zhu, F. 2012. A survey of recommender systems in twitter. In *Social Informatics*, volume 7710 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg. 420–433.

Li, Q.; Wang, J.; Chen, Y. P.; and Lin, Z. 2010. User comments for news recommendation in forum-based social media. *Inf. Sci.* 180(24):4929–4939.

Lu, C.; Lam, W.; and Zhang, Y. 2009. Twitter user modeling and tweets recommendation based on wikipedia concept graph. Technical Report WS-12-09, AAAI Technical Report.

Marquardt, J.; Farnadi, G.; Vasudevan, G.; Moens, M.-F.; Davalos, S.; Teredesai, A.; and De Cock, M. 2014. Age and gender identification in social media. In *Proceedings of CLEF 2014 Evaluation Labs, CLEF 2014 Evaluation Labs, Sheffield, UK, 15-18 September, 2014*. Accepted.

Moro, A.; Raganato, A.; and Navigli, R. 2014. Entity Linking meets Word Sense Disambiguation: a Unified Approach. *Transactions of the Association for Computational Linguistics (TACL)* 2:231–244.

Myers, S. A., and Leskovec, J. 2014. The bursty dynamics of the twitter information network. *CoRR* abs/1403.2732.

Navigli, R., and Ponzetto, S. P. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence* 193:217–250.

Sap, M.; Park, G.; Eichstaedt, J.; Kern, M.; Ungar, L.; and Schwartz, H. A. 2014. Developing age and gender predictive lexica over social media. In *EMNLP-2014: the 2014 Conference on Empirical Methods in Natural Language Processing*.

Siehndel, P., and Kawase, R. 2012. Twikime! - user profiles that make sense. In *International Semantic Web Conference (Posters & Demos)*.

Smith, J. 2014. Gender prediction in social media. *CoRR* abs/1407.2147.

Stilo, G., and Velardi, P. 2014. Time makes sense: Event Discovery in Twitter using Temporal Similarity. In *IEEE/WIC/ACM International Conference on Web Intelligence (WI-'14) (To appear)*.

Szell, M., and Thurner, S. 2013. How women organize social networks different from men: gender-specific behavior in large-scale social networks. *Scientific Reports* 3:1214.

Tarjan, R. 1972. Depth-first search and linear graph algorithms. *SIAM Journal on Computing* 1:146–160.

Tiernan, J. C. 1970. An efficient search algorithm to find the elementary circuits of a graph. *Commun. ACM* 13(12):722–726.

Wagner, C.; Liao, V.; Pirolli, P.; Nelson, L.; and Strohmaier, M. 2012. It's not in their tweets: Modeling topical expertise of twitter users. In *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Conference on Social Computing (SocialCom)*, 91–100.