

Mapping communities in large virtual social networks

Using Twitter data to find the Indie Mac community

Michiel van Meeteren, Ate Poorthuis, Elenna Dugundji

Department of Geography, Planning and International Development Studies

Universiteit van Amsterdam, Netherlands

michielvanmeeteren@gmail.com, atepoorthuis@gmail.com, e.r.dugundji@gmail.com

Abstract— This paper describes a multi-method approach to delineate a “real world” community of practice from a large N dataset derived from the social networking site Twitter. The starting point is previous qualitative research of a virtual community of independent (“indie”) developers who create software for Apple’s Macintosh and iPhone platforms. Indie developers have been active on Twitter from an early stage on and they use Twitter to sustain interactions between peers, exchange technical information and for viral “echo chamber” marketing. The publicly available Twitter API is used to mine a network consisting of several million edges, which is sized down to a large network containing roughly 1 million edges through several pruning methods. The fast greedy algorithm is then used to detect subgraphs within this large network. Triangulation with qualitative data proves that the fast greedy algorithm is able to distill meaningful communities from a large, noisy and ill-delineated network. The accuracy of this approach gives rise to the discussion of the value for businesses and market research, since it offers opportunities to identify and monitor target audiences at a finely grained level. However, we should be wary of the serious consequences with regard to privacy and ethics. The proposed multi-method approach allows micro level inferences from a macro dataset of which the individual Twitter user might be completely unaware. The results could have consequences for the anonymity of key persons behind the scenes of social and political movements or any other communities whose members are active on Twitter or other social networks.

Keywords- social networking software; Twitter; indie developer entrepreneur; social network analysis; community detection

I. INTRODUCTION

Social networking software is increasingly becoming an indispensable feature of our social lives. It influences how we go about with our friends, how we sustain contacts while being geographically separated and it allows businesses in some sectors to flourish, because they are relieved of the friction of distance. The resulting large virtual social networks are both a meaningful social entity - since they shape social practices - and an interesting data source for academics. However, the barriers of entry to online networks are generally low, introducing noise in the resulting online data. To be able to make any inferences on the level of meaningful actual social relationships, we need to remove noise from any data that is mined from online resources. This paper will address this problem and propose a method of de-noising large N networks as obtained from the social software platform Twitter¹. The paper assesses the network structure for a community of early

Twitter adopters, the Indie Mac developer community. We build upon qualitative research that was conducted previously on this community [1,2] and use those qualitative findings to assess the validity of our proposed method to infer socially meaningful communities from ill-defined, noisy large N datasets. This paper intends to use social network analysis for the following two aims: 1) isolate and map the Indie Mac community that was studied in *Indie Fever* [1] from an ill-delimited noisy large N virtual network mined from Twitter; and 2) describe the social structure of this community by analyzing a variety of network measures.

The paper is built up as follows: Section two is a brief overview of some existing literature on community detection. Sections three and four provide the necessary background on the Indie community, their use of social software and the role of certain key figures - or tastemakers - in the community. In section five we lay out the method that we used to infer socially meaningful communities from a large, noisy online social network. Section six presents results of the network mapping and will corroborate network characteristics of the total network and its discernible sub-communities. Section seven elaborates on the internal attributes of the Indie community by analyzing the network measures of that subgraph. Section eight discusses issues of validity, the potential business value, social and legal issues, and potential moral hazards of the described method. Section nine draws some conclusions and sketches avenues for further research.

II. LITERATURE REVIEW

The theoretical and technical advancement of a structural definition of the concept of a group has been a realm of the fascination of social scientists since Simmel’s landmark work at the turn of the last century. [3] Early contributions to understanding substructure in networks took a “bottom-up” approach, starting from individual members and considering their connections to others, building up from micro to macro structures. [4] In 1949, Luce and Perry introduced the concept of a *clique*, where all members are directly connected to all other members, otherwise known as a complete subgraph. [5] A series of variations on this concept followed in the subsequent decades, relaxing the strict distance requirements: *n-clique* [6], *n-clan* [7], *n-club* [8]; and alternatively the strict density requirement: *component* [9], *k-plex* [10], *k-core* [11].

In 1983, Seidman suggested that the electrical engineering concept of an *LS-set* might provide a useful network formalization of the sociological notion of a group. [12,13] An *LS-set* is defined such that each of its proper subsets has more

¹ <http://twitter.com>

connections to its complement within the group, than to other actors outside in the larger network. In 1990, Borgatti, Everett and Shirey relaxed the strict LS-set in various directions. [14] They introduced a 2-step algorithm for computing *lambda sets* based on evaluation each of the connections in a network in terms of how much of the total *flow* among actors in the network go through the connection [15,16], then applying hierarchical clustering [17]. This approach differs from the above techniques to studying groups, in starting from the perspective of the entire network, proceeding from macro structure to micro structure.

In 2002, physicists Girvan and Newman proposed a landmark “top-down” approach to finding substructure in networks based on the measure *edge betweenness*, a generalization of Freeman’s node betweenness centrality [18,19]. Their algorithm evaluates each of the connections in the network in terms of the number of shortest paths between pairs of actors that run along the connection [20], removes the connection with highest betweenness, and iterates. Radicchi et al applied a local approach based on the measure, *edge clustering*, a generalization of node-clustering. [21] To facilitate the object discrimination of meaningful groups, they also introduced definitions of strong and weak communities, similar, respectively, to LS-sets and k-cores.

In 2004, Newman and Girvan introduced the concept of *modularity*, calculating the fraction of inward connections within a group minus the expected fraction in an equivalent network with the same community divisions but with random connections between actors. [22] Modularity can be either positive or negative, with positive values indicating the possible presence of community structure. In a series of advances that rapidly followed in the next two years, physicists introduced various modularity optimization methods for dividing networks using techniques such as: simulated annealing [23,24,25], short random walks [26], fast greedy algorithms [27,28], extremal optimization [29], and spectral graph partitioning combined with fine-tuned search [30,31].

Several of the modularity optimization algorithms have been implemented in *igraph* [32], an open source network analysis package in R [33]. Since the fast greedy algorithm is the only one which is capable of handling networks of millions of nodes, we use this further in our analysis. However, it is important to note that our problem is different from a classical community detection problem in an important way. In our case, the starting network from which we would like to recover the community is *ill-defined*. It would be infeasible given the computational means available to the authors to mine the entire Twitter network of all users. We proceed therefore in an iterative way, guided by qualitative knowledge about the network under study, using the fast greedy optimization of modularity together with other social network analysis techniques as part of a multi-method approach.

III. THE INDIE MAC COMMUNITY

The “Indie Mac” developer community refers to a group of independent software companies that develop software for Apple’s Macintosh platform. The majority of them are one-person shops, except for the more successful ones who sometimes have a few employees (although more than ten is

exceptionally rare). These companies sell their software to worldwide markets over the Internet, circumventing the traditional costs of physical production and distribution, which require substantial capital investments. Despite the fact that these software companies could regard each other as competitors, there is a lively interaction between them. This is done primarily through online means since they rarely are physically co-located. In time, a specific Indie culture and habitus has developed among them that guides interaction and informal social hierarchy between Indie developers [2]. They could be considered a virtual community of practice [34] where the specific habitus and related tacit knowledge embedded within it guides the ideas about how Mac software should look, feel and function.

Since the introduction of the iPhone developer platform to 3rd party software developers in March 2008, the number of developers working with Apple’s Cocoa software development technologies (Mac and iPhone software development technologies are strongly related) has increased enormously. Some Indie developers that were traditionally working on the Mac platform have also developed iPhone software. Previous research has shown that although the vast majority of iPhone developers does not identify with the traditional Mac developer community, there is a core that does. These developers get integrated in the wider Apple developer community through interaction on blogs and on events like Apple’s annual worldwide developer conference [35].

IV. THE ROLE OF ONLINE COMMUNICATION AND VIRTUAL SOCIAL NETWORKS

Social software can be considered the infrastructural backbone of the Indie community. A variety of one-to-one, one-to-many, and many-to-many social software was in use at the time of previous qualitative fieldwork during February-March 2008. Blogs, mailing lists, IRC chats, wikis, chat-clients and Twitter (which was steeply on the rise at the time) were reported to be frequently used. When analyzed functionally, the online behavior of Indie developers broadly performs three functions: identification and socialization, satisfactions of informational needs, and marketing [1]. In this section, these functions will be highlighted, and the importance of Twitter and the role of tastemakers will be explained.

A. Identification and socialization

The plethora of blogs, mailing lists, and online chatter together foster an “Indie identity,” “sense of belonging” and “profession-specific culture” that is characteristic of communities of practice. It disseminates the tacit knowledge² of how “good Indie” software should look, feel and function and also how proper behavior amongst developers should be conducted. In addition to the quality of software, “proper” online behavior played a big role in which Indies are held in high esteem by their peers [2]. The symbolic capital [36] within the Indie community gained in this way can be utilized to enhance “echo chamber” marketing, get access to other developers and probably in gaining access to Apple Inc.

² Reference [1], pp. 51-52 elaborates why this online knowledge should still be considered “tacit” despite the fact that the sources of that knowledge are ubiquitous.

B. Satisfaction of information needs

A second function of the online behavior is that it helped the spatially dispersed Indies to diffuse information. The online networks work as a “virtual watercooler” [37]. They help in providing peer camaraderie and friendship while coding software, which can be a long solitary task. In addition, it allowed Indies to disseminate industry rumors, get help on coding problems and find references on people applying for contracting work.

C. Marketing

Lastly, the online network of Indie developers plays a role in the marketing of their software. The Indie market can - because of its lack of physical production and distribution costs - be considered a “long tail” market [38]. The product is available to everyone with an Internet connection, but people need to find it and appreciate it in order to buy it. This means that online exposure is the crucial factor in marketing and that the prime determinant of economic success is derived from getting your software known past a certain “tipping point” [39]. Because the signal-to-noise ratio is very high on the Internet, there is a high potential added value of peer recommendation of software products. Indies use echo marketing as a form of peer review. If a new software title is released, other developers endorse it if they appreciate it - and often only if they jointly appreciate the software title and the developer who made it. These endorsements go through the online network, often reaching the specialized journalists of the Macintosh world. Thus, the size and structure of the online network and the inclination of other developers and intermediaries to “echo” the message influence the economic success of a developer to a very high degree [2].

D. Twitter

Twitter has grown by 742% in 2008³ and is at the moment adopted worldwide as the latest web 2.0 innovation to reach mass audiences, politicians and celebrities. However, within the Indie Mac community, Twitter gained a critical mass of users relatively early and its use was omnipresent at the time of fieldwork and still is today. One of the reasons for this fast adoption is that a well-known Indie developer, Craig Hockenberry of the Iconfactory, developed a desktop client for Twitter early on, which let Twitter run in the background so the user could concentrate on other things. The idea behind Twitter is that you can post messages with a maximum of 140 characters on the Internet, which subsequently can be read by everybody who is “following” you. You only get the messages from those whom you follow. This allows you to simultaneously broadcast a message to a lot of people while being able to limit the amount of information that reaches you. After the advent and take off of the iPhone, a lot of Twitter activity has moved to that platform. A variety of Twitter clients is available for the iPhone and there is a fast paced but friendly-voiced competition on innovation going on between them (Kyle Baxter, 07-05-2009)⁴.

³ <http://mashable.com/2009/01/09/twitter-growth-2008/>

⁴ <http://www.tightwind.net/2009/05/another-example-of-the-greatness-that-is-the-mac-community>

E. The role of tastemakers

Within such a business environment, it is evident that social capital plays a role in the economic chances of a company. Having your software endorsed by the community can be a great asset in economic terms. This endorsement usually follows from complying with the aesthetic and social discourses that guide “proper” behavior in the community. The literature on the cultural industries has emphasized the role of tastemaking actors in this respect [40]. A tastemaker could be defined as an intermediary actor who yields -often symbolic- power and uses that power, by for example endorsements, to help selected authors to become economically successful. Throughout history, journalists, art critics and famous artists have played important roles in making new talent famous: such as can be found in the music industry or the French literary field in the 19th century [40,41]. Respondents acknowledged that these “tastemakers” play an important role in the Indie Mac community, especially because these persons often function as a bridge between the in-group of developers and the first tier of critical users.

The person who recurrently arose as an important tastemaker is technology journalist John Gruber, who maintains the Daring Fireball blog.⁵ Gruber, who has an educational background in computer science, is considered to be an important software “connoisseur” and tech industry insider by both Indie developers and the wider audience. The Daring Fireball RSS feed has over 150,000 subscribers and the website has an estimated number of 1.3 million page views per week.⁶ The tastemaker role is also signified by his Twitter statistics. On May 16, 2009 he had 27,191 followers while following only 311 people - most of who were reciprocal.

V. POSSIBILITIES OF QUANTIFYING QUALITATIVE ANALYSIS VIA SOCIAL NETWORK ANALYSIS

One of the problems when researching a virtual community of practice without registered membership is that you lack the means of delimitating the population. Membership is a matter of degree, and boundaries are blurred between actual developers, contributing enthusiastic users [42], journalists and other secondary agents and noise. Despite these difficulties, the qualitative evidence infers that within the community there is a shared habitus, and culture comparable to what usually is found within more formal organizational boundaries [43]. In this section, we set out to detect this community within the noisy, large network that Twitter is. Since John Gruber was pinpointed by Indies as a prime example of an important tastemaker, it can be assumed that most developers who are part of the Indie field as described in [1] and who use Twitter can be found within one degree of separation of John Gruber’s Twitter network.

Once the specific Indie community is located we can use network centrality measures to analyze the internal stratification of the community of Indie developers. This will be elaborated in section seven.

⁵ <http://daringfireball.net>

⁶ <http://daringfireball.net/feeds/sponsors/>

A. First Stage Data Collection

The Twitter network is a directed network. Users can follow other users but this link does not have to be reciprocal. In Twitter-speak, every user has “followers” (out-degrees) and “friends” (in-degrees).⁷

John Gruber was originally used as the starting point of the network. Every user that is within 2 degrees from Gruber is included. In practice, that means everyone who Gruber is “following” (approx. 300 users) or is following him (27,000 users) and all users that are following them or are friends of those users. For every node in the network, we will store metadata including name, location, description, date of joining, source of tweets (Twitter client) and total number of tweets posted.

To collect data from Twitter, we use a combination of techniques: the Twitter API service,⁸ Ruby to script the mining part, and MySQL to store the data. Twitter provides an extensive Search and REST API, from which the “users/show,” “friends/ids” and “followers/ids” methods are used. We will interface with the Twitter API using the Ruby gem Grackley. A Ruby script is written that: 1) harvests all friends and followers using “friends/ids” and “followers/ids” methods from the API; 2) collects metadata of these nodes; 3) recursively harvests all friends and followers of those nodes; 4) collects metadata of those nodes; and 5) stores all data in a MySQL database.

B. Pruning

This first stage of data collection results in an enormous network with 28 million edges and 4 million vertices. We now cut off the network at 1 degree of separation from Gruber, but we keep all the edges between alters that fall within this condition. We now have an ego-network with 27,218 nodes and 1.5 million edges.

Albeit considerably smaller, this network still contains a lot of noise caused by institutional Twitter accounts and other accounts that are not maintained by a single individual. We argue that, for any individual, it is not possible to follow more than 600 people. Following such an amount of Twitter users simply results in a complete information overload. As real, individual, users use Twitter exactly for communication and information purposes, excluding all users that follow more than 600 people is a sound way to de-noise the network. Eliminating those users produces a considerably smaller network of approximately 22,000 nodes and 440,000 edges.

C. Community Detection

For the analysis of this network, we will use a combination of the open source statistical software R and one of its social network analysis libraries, igraph. To pinpoint the Indie developer community, an edge list is loaded into R. To detect the indie community, we use the fast greedy algorithm developed by Clauset, Newman and Moore (CNM). It was specifically designed to detect community structure in very large networks and has a relatively low computational cost. It

was originally tested on the Amazon.com recommendation network, in which edges are drawn between two objects in Amazon’s store if multiple customers buy both. The algorithm found meaningful communities within this network, delineating them into categories such as jazz, engineering, children’s video’s, etc. [28] In our case, the fast greedy algorithm finds 5 dense subgraphs within the total network, containing respectively 5577, 8780, 445, 6172 and 780 nodes.

D. Triangulation with Qualitative Field Data

Triangulation with qualitative fieldwork data [2] reveals that all interviewees from the previous research can be found in subgraph 4 (6172 nodes). This gives enough reason to believe that subgraph 4 contains the Indie developers cluster. Network centrality measures (in-degree and out-degree centrality, closeness centrality, betweenness centrality and eigenvector centrality) are thus calculated for this subgraph.⁹

From the analysis of central nodes, it becomes clear that, although Gruber’s ego network gives a remarkably complete picture of the Indie community, there are some important parts missing. Therefore, we will introduce additional egos and mine the same data for them as we did for Gruber.

E. Second Stage Data Collection

Among the additional egos are the 15 most central (eigenvector centrality) nodes in the detected community and 7 additional egos based on qualitative findings from fieldwork data. As a result there are now 23 egos (including John Gruber) that are used as a new starting point for data collection repeating the same procedure described for the first stage.

F. Re-pruning

We then apply the same pruning procedure again on this extended network to bring it back down to manageable and meaningful proportions. Only nodes are selected that are within 1 degree of separation from the egos in the aforementioned list. This results in a network of approximately 52,000 nodes.

Then, we select only those nodes that follow less than 600 other users, based on the same assumptions as above. This results in a final network that we will use for further analysis and consists of 40,512 vertices and 1,023,317 edges.

G. Community Re-detection

We run the fast greedy algorithm on this network again to get a more accurate pinpointing of the Indie developer community than based on Gruber’s ego network alone. The algorithm again finds five communities, respectively with 13978, 11522, 591, 1428 and 12669 nodes. All Indies are now in subgraph five (N=12669), which is subsequently called “the Indie community.”

The following section will examine the full network and the Indie community subgraph and establish some qualitative differences between the various subgraphs based on the “description” field of each user. This description is provided by the user and often hints at the occupation a user has.

⁷ Whether followers and friends should be considered out- or in-degree is debatable; we chose to follow the direction of information flows and these followers are out-degree.

⁸ <http://apiwiki.twitter.com/>

⁹ These detailed tables are not included here due to space limitations as well as potential privacy issues and legal concerns.

VI. COMPARING THE TOTAL NETWORK AND THE DISTINCTIVE INDIE NETWORK

As has been noted, the final network is composed out of five distinct communities that were detected by the CNM fast greedy algorithm. We interpret the different subcommunities by using frequency tables of words in the “description” field to see whether the algorithm does indeed provide qualitatively meaningful distinctions. For easier analysis, we convert the frequency tables to word clouds, which is nothing more than a graphical representation of the frequency of words. It plots words with a higher frequency larger than those words that recurring less. In our case, it gives an insight into the essence of each community with the glimpse of an eye. For this paper, we use the service provided by Wordle¹⁰ since it allows filtering out common words in the English language to get better results. The description fields of all nodes were aggregated into a large text file, which then was fed into the Wordle service.

The different word clouds for the total network and the various subgraphs are visualized.¹¹ The word cloud of the total network shows a reflection of the interests of the readers of Gruber’s blog: we can clearly discern the various creative industry and technology related keywords that we would expect of a creative technology guru. Keywords like photography, design, music, web, technology, media, geek¹² and developer stand out (Fig. 1).



Figure 1. Word clouds of total network and subgraph five.

¹⁰ <http://www.wordle.com>

¹¹ Due to space limitations, only the word clouds for the total network and subgraph five are reproduced in this paper. For full color versions of the total network and all 5 subgraphs the interested reader is referred to: <http://www.humangeographer.com/wordcloud.html>

¹² Geek is a honorary nickname for technology enthusiasts.

Contrastingly, community five - the Indies - seems to quite neatly capture the interests of software developers. Keywords like: Mac, developer, iPhone, software and Apple stand out. This is already a qualitative reflection that the community detection in section five - which does not use keyword data, but solely relies on edges - does quite a good job at isolating the relevant community in the whole network. Only community three, which has only 591 members seems to have a relevant overlap but is in general more varied than community five.

Community two is interesting because it seems to capture another specific field of the Gruber audience: web, development, and design dominate this subgraph. These keywords indicate an interest towards topics like web development and design, which is a distinct field from Mac/iPhone software development. Lastly, communities one and four show a more varied collection of keywords. They capture other creative fields like music, photography and media and show a higher degree of “consumer” keywords like fan, student, geek and enthusiast.

VII. ANALYZING THE INDIE COMMUNITY

A. Network Centrality Measures

This section will provide a data analysis on the community five, the Indie community. For all the nodes in the community (N=12699), the indegree and outdegree centrality, closeness centrality, betweenness centrality and eigenvector centrality are calculated. Subsequently all the nodes were ranked according to these five centralities, with the highest centrality receiving position one, the second highest position two, etc.

In the context of the Indie community, the rank according to the different centralities has the following meaning:

1) *In-degree centrality*: The number of other actors in the developer network (community five in respect to the total Gruber network) which a given actor is following (i.e. receives information from).

2) *Out-degree centrality*: The number of other actors in the developer network community that follow a given actor (i.e. that a given actor sends information to). Nodes (actors) with a high outdegree centrality ranking tend to be very popular within the Indie community.

3) *Closeness centrality*: Closeness centrality is a measure to indicate the total distance of a node to the other nodes in the network [19]. In this paper we use an *inverted* measure, meaning that the highest scoring node is relatively closest to all other nodes in the network.

4) *Betweenness centrality*: Betweenness centrality indicates how often a node falls in the shortest path between two other nodes. [19]

5) *Eigenvector centrality*: Eigenvector centrality is a measure designed to measure “power” positions within a social network [44]. Nodes with a high score on the eigenvector ranking are connected to other nodes with a high eigenvector centrality. Thus, eigenvector centrality is a potent tool to identify the “core” of a community, in terms of network connectivity. Eigenvector centrality will be used as the primary ranking order in this section because it is the most

potent way to approximate the status of an Indie developer relative to the peer group. Moreover, a qualitative cross check with a subject in the Indie community acknowledged that the ranking by eigenvector centralities was a very plausible and meaningful ranking from the perspective of an Indie developer.

B. Occupation

Apart from the centralities, we qualitatively inferred the occupation of the node for the twenty highest ranking nodes in each ranking. Seven different “occupations” were established, of which three are further discussed in this analysis: 1) “Developer,” a generic category of software engineers and developers who do not work for Apple; 2) “Apple,” people currently employed by Apple Inc.; and 3) “Journalist,” freelance technology journalists, maintainers of popular blogs, or employees of (online) magazines in the technology press.

C. Discussion

For a network as large as the Indie network we could argue that both betweenness and closeness centralities are measures that say something about the position of a node relative to all the nodes in the network, whereas eigenvector centrality says something about the position of the node in relation to other important nodes. It should be noted that all three measures assume a symmetric network which implies that it is theoretically possible (by following the right people) to achieve a high ranking in the network without ever having these ties reciprocally acknowledged. In-degree and out-degree centrality are asymmetrical measures and can therefore be used to control for that problem. Actors which have a high “self appointed” eigenvector centrality will also show a high ranking on in-degree centrality but a relatively low ranking on out-degree centrality (a given actor follows a lot of people, explaining the high in-degree ranking, but not a lot of people follow him or her, hence the low out-degree ranking).

The top twenty nodes of each centrality measure are described in tables; they contain the actual value, the ranking, name, and the occupation of the node.¹³ To be able to compare the values, the eigenvector centrality ranking is included as well. When we look at in-degree centralities, it is interesting to see that apart from nodes that were used as a starting point, all the other nodes show a relatively low score on the eigenvector ranking. Further studying would be required to see whether these developers are either “aspiring developers” that attempt to make a name in the community by deliberately following the well known “Indie stars” or that they are part of subcommunities within the Indie network that show less clustering than the top tier. The top twenty out-degree centralities could be considered as the result of a “popularity” contest among the whole network - most people in the network choose to follow these nodes. This explains the high ranking of some institutional accounts¹⁴ and journalists. Also, “visible”

developers are in this list. However, this does not necessarily mean that the out-degree mirrors the eigenvector values. Institutional accounts usually score lower in terms of eigenvector centralities, as do journalists.

When we compare the rankings for closeness and betweenness centrality, two differences stand out. Firstly, it appears that nodes with a high betweenness rank often have a far lower eigenvector rank than you would intuitively expect - ranging far in the double or even triple digits. It seems that being a spider in the network does not imply that you are automatically part of the in-crowd of top ranking eigenvector centralities. A second interesting finding is that the journalists are mostly absent from the betweenness ranking while they prominently feature in the closeness ranking. Apparently, journalists are not so much “connectors” in the network but from this “relative periphery” they are able to gather information from the network as a whole which intuitively complies with their job description.

Three tables compare the centrality rankings in the Indie community for the three most relevant occupations: developers, journalists, and Apple employees (Tables I-III).¹⁵

When we examine the developers we see that the highest ranking developers in terms of eigenvector centralities generally have high out-degree, closeness and betweenness centrality rankings as well.

TABLE I. CENTRALITY RANKINGS OF THE HIGHEST RANKING DEVELOPERS IN THE INDIE COMMUNITY ^a.

Eigenvector	In-degree	Out-degree	Closeness	Betweenness
1*	114	8	4	4
2*	17	13	8	2
3*	78	12	7	8
4*	98	1	1	1
5*	1138	4	3	13
6*	2	20	11	3
7*	962	6	2	6
8*	531	11	9	15
9*	5	36	26	9
10*	196	21	17	42
11*	74	25	21	11
13	107	1007	19	23
14*	1915	1707	15	38
15	77	523	61	95
16	518	1355	14	39

a. based on eigenvector centrality ranking

TABLE II. CENTRALITY RANKINGS OF THE HIGHEST RANKING APPLE EMPLOYEES IN THE INDIE COMMUNITY ^a.

Eigenvector	In-degree	Out-degree	Closeness	Betweenness
12*	208	22	20	29
17	105	37	42	41
20	171	48	23	96
21	66	42	40	30
22	42	80	28	31
31	590	43	141	211
35	1130	41	82	238

a. based on eigenvector centrality ranking

¹³ These detailed tables are not included here due to space limitations as well as potential privacy issues and legal concerns.

¹⁴ Institutional nodes are dedicated Twitterfeeds to spread information according to a certain theme (eg. a specific magazine or a conference). A significant number of these accounts were filtered out when we excluded all nodes with more than 600 friends, but some remain.

¹⁵ Nodes marked with an asterisk are egos used for data mining which could bias the results in terms of in-degree and out-degree ranking.

TABLE III. CENTRALITY RANKINGS OF THE HIGHEST RANKING JOURNALISTS IN THE INDIE COMMUNITY^a

Eigenvector	In-degree	Out-degree	Closeness	Betweenness
19	2054	15	13	70
25	1351	9	10	20
27	495	16	16	37
29	101	49	108	27
53	419	45	371	53

a. based on eigenvector centrality ranking

Interestingly enough, the last 4 developers do not fit that pattern. One of the reasons is a bias towards those egos who were used as a starting point for data collection, which explains higher in-degree and out-degree centralities.¹⁶ However, these developers could also be considered more “in-crowd” developers: people who have a high status but are not as publicly known as some others.

This picture is strengthened when we look at Apple employees with high eigenvector centrality rates; they also show lower scores relative to their eigenvector score. The rationale behind this is that Apple developers are supposedly less inclined to become “famous” in the Indie community. However, knowing Apple employees can be a very important knowledge asset for Indie developers. Therefore it seems logical that an in-crowd of A-list developers in the network who know who to look for have stronger connections to Apple employees than average members of the community. This could explain their higher eigenvector centrality relative to other centrality measures.

Lastly, when we look at journalists the aforementioned relationship between closeness and betweenness seems true for the highest three ranking journalists but not for numbers four and five. However the latter two seem more actively inclined to follow Indie developers, as is indicated by the lower rankings in terms of in-degree centrality

VIII. IMPLICATIONS

A. Validity

The results of the word cloud analysis and the comparison of centralities show that it is possible to meaningfully isolate subcommunities from a large N network. This implies that the fast greedy algorithm which was originally tested on the Amazon.com recommendation network is also capable of detecting different social groups in real-world social networks. Another important issue when we want to draw inferences from large N social networks is to gauge the extent in which online ties reflect “real world” social ties. Anyone on Twitter can strategically befriend themselves to important key persons without having a “real world” connection. This means that at first glance people which score high on centrality measures do not necessarily also have an important position in the social network outside of Twitter. The above analysis shows that using filtering techniques and by comparing centralities inferences can be drawn about which links in the network reflect meaningful social relationships and which do not; it

reveals the “hidden social network” [45]. Further testing on different empirical cases could provide insights to standardize this analytical procedure.

B. Business and marketing

The word clouds of the various subgraphs in the Gruber network show meaningful cohesion in terms of keywords prevalent in the subgraphs. This opens up opportunities for business both in terms of market research and direct marketing. The methods applied in this paper can find subcommunities that share common interests within a larger Twitter network. These communities form finely grained target audiences for marketing purposes at almost no cost, which can be very valuable for targeted advertising. Moreover, this identification of communities can be of great use for further market research. The possibilities are numerous: we can track changes in the composition of the communities; we can gauge what is going within communities by analyzing their Twitter messages; we can even check the effectiveness of marketing by looking at how each community tweets about our brand. We can do this all without paying for panels or having the users know they are being observed.

C. Consumers, privacy and ethical concerns

These last remarks on advertising touch upon potential ethical issues that are involved with this kind of network analysis. The core ethical issue is that the method allows one to make inferences on the individual or small community level by using massive datasets. The resulting individual level data is information of the relation between an individual and a wider social structure of which the individual is not necessarily aware. This raises issues on the lack of informed consent on the individual. We can conduct the aforementioned market research with public information, without the subjects’ explicit awareness. Interestingly, Twitter hardly mentions anything on privacy and it is very easy to extract data from Twitter, even as a non-authenticated user (as opposed to most other online social networks). In fact, the only clause in the Terms of Service that mentions anything about data use by third parties, encourages it.

D. Political issues and moral hazards

Lastly, there is an implicit moral hazard of a larger magnitude. Twitter is increasingly being used as a political mobilization instrument. For example, during the alleged election fraud in Iran in the summer of 2009, Twitter was used to coordinate demonstration. In this case the Iranian government cracked down the Twitter activism by blocking access to Twitter. However, if a government uses methods like the one described in this paper, it would be able to identify the central nodes in a movement even if these are not visible to the wider public. This could subsequently allow someone with malicious intent not only to cripple a social organization by eliminating the nodes in the network that perform crucial behind-the-scenes roles, but this would also put people in general physical danger. At this moment, this analysis is possible for anyone with the knowledge to access the Twitter API – or the API of most other online social networks .

¹⁶ Some of this bias has been filtered out in the community re-detection process. However, the presence of friends and family is unavoidable without determining a cutoff point in the network.

IX. CONCLUSIONS AND AVENUES FOR FURTHER RESEARCH

In this paper we have limited ourselves to a small analysis of only part of the available data. However, by triangulating with earlier qualitative findings and with a self-description provided by each node in the network, we have shown that it is possible to find meaningful subcommunities in a large, noisy online social network using the multi-method approach presented here. Community detection has been widely applied in recent years to diverse networks using various algorithms [46,47]. Our work differs from a classical community detection problem in that starting network from which we would like to recover the community is ill-defined. We proceed in an iterative way, using the fast greedy optimization of modularity together with other social network analysis techniques as part of a multi-method approach during data mining.

The ease of accessing this public data, combined with the relatively low technical requirements and the accuracy of community detection algorithms, makes the mining of online social networks and subsequent community detection a powerful tool for both academic and market research. However, these same elements also open the road to less benign uses and oblige us to think about the consequences of making data public.

This specific analysis provided interesting insights in the community of Indie developers, which would otherwise have been difficult to research quantitatively due to lack of a meaningful sampling frame. By using the Twitter network as a proxy for actual social relations, we can infer quantitative insights on the internal stratification within a virtual community. The selective connections of journalists, Apple employees and the discovered internal stratification patterns reaffirm and expand earlier qualitative research. However, we should keep in mind that Twitter connections are a mere proxy for the social relations that these virtual relations reflect. There could be other actors involved who do not actively use Twitter, and actors who use Twitter strategically could end up higher in the rankings than their actual status within a community is.

Until now, we have not even delved into the content of the actual tweets. It is possible to explore actual communication between users, look at the dispersion of re-tweets (literal quoting of other users), and study the evolvement of choice behavior over time [48]. With these sorts of advanced techniques, even more detailed analysis of the dynamic network structure of communities could be performed.

REFERENCES

- [1] M. van Meeteren, Indie fever: The genesis, culture and economy of a community of independent software developers on the Macintosh OS X platform, 2008. <http://www.madebysofa.com/indiefever>
- [2] M. van Meeteren, working paper, 2009. <http://www.humangeographer.com/pdf/ethnographyofIndiesWP.pdf>
- [3] G. Simmel, *Am. J. Soc.*, vol. 8, no. 1, pp. 1-46, 1902.
- [4] R. A. Hanneman and M. Riddle, *Introduction to Social Network Methods*. Riverside: Univ. of California, Riverside, 2005. <http://faculty.ucr.edu/~hanneman/>
- [5] R. D. Luce and A. Perry, *Psychometrika*, vol. 14, pp. 94-116, 1949.
- [6] R. D. Luce, *Psychometrika*, vol. 15, pp. 169-190, 1950.
- [7] R. D. Alba, *J. Math. Soc.*, vol. 3, pp. 113-126, 1973.
- [8] R. J. Mokken, *Quality and Quantity*, vol. 13, pp. 161-173, 1979.
- [9] F. Harary, R. Z. Norman and D. Cartwright. *Structural Models: An Introduction to the Theory of Directed Graphs*. New York: John Wiley and Sons, 1965.
- [10] S. B. Seidman and B. L. Foster, *J. Math. Soc.*, vol. 6, pp. 139-154, 1978.
- [11] S. B. Seidman, *Social Networks*, vol. 5, pp. 269-287, 1983.
- [12] S. B. Seidman, *Social Networks*, vol. 5, pp. 92-96, 1983
- [13] F. Luccio and M. Sami, *IEEE Transactions on Circuit Theory*, vol. CT-161, pp. 184-188, 1969.
- [14] S. P. Borgatti, M. G. Everett and P. R. Shirey, *Social Networks*, vol. 12, pp. 337-358, 1990.
- [15] L. R. Ford Jr. and D. R. Fulkerson, *Can. Journal of Math.*, vol. 8, pp. 399-404, 1956.
- [16] R. E. Gomory and T. C. Hu, *Journal of SIAM (Appl. Math.)*, vol. 12, p. 348, 1964.
- [17] S. C. Johnson, *Psychometrika*, vol. 32, pp. 241-254, 1967.
- [18] M. Girvan and M. E. J. Newman, *Proc. Natl. Acad. Sci. USA*, vol. 99, pp. 7821-7826, 2002.
- [19] L.C. Freeman, *Social Networks*, vol. 1, no. 3, pp. 215-239, 1979.
- [20] M. E. J. Newman, *Proc. Natl. Acad. Sci.*, vol. 98, pp. 404-409, 2001.
- [21] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, D. and Parisi, *Proc. Natl. Acad. Sci. USA*, vol. 101, pp. 2658-2663, 2004.
- [22] M. E. J. Newman and M. Girvan, *Phys. Rev. E*, vol. 69, 026113, 2004.
- [23] R. Guimera, M. Sales and L. A. N. Amaral, *Phys. Rev. E*, vol. 70, 025101, 2004.
- [24] R. Guimera and L. A. N. Amaral, *Nature*, vol. 433, pp. 895-900, 2005.
- [25] J. Reichardt and S. Bornholdt, *Phys. Rev. E*, vol. 74, 016110, 2006.
- [26] P. Pons and M. Latapy, *J. Graph Alg. and Appl.*, vol. 10, p. 191, 2006.
- [27] M. E. J. Newman, *Phys. Rev. E*, vol. 69, 066133, 2004.
- [28] A. Clauset, M. E. J. Newman, and C. Moore, *Phys. Rev. E*, vol. 70, 066111, 2004.
- [29] J. Duch and A. Arenas, *Phys. Rev. E*, vol. 72, 027104, 2005.
- [30] M. E. J. Newman, *Proc. Natl. Acad. Sci.*, vol. 103, pp. 8577-8582, 2006.
- [31] J. H. Ruan and W. X. Zhang, *Phys. Rev. E*, vol. 77, 016104, 2008.
- [32] G. Csardi and T. Nepusz, *InterJournal, Complex Systems* 1695, 2006. <http://igraph.sf.net>
- [33] R Development Core Team. *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing, 2009. <http://www.R-project.org>
- [34] A. Amin and J. Roberts, *Research policy*, vol. 37, no. 2, pp. 353-369, 2008.
- [35] M. van Meeteren, working paper, 2009. <http://www.humangeographer.com/pdf/valuechaincomparisonWP.pdf>
- [36] P. Bourdieu, *The Field of Cultural Production*. Cambridge: Polity, 1993.
- [37] J. S. Brown and P. Deguid, *The Social Life of Information*. Boston: Harvard Business School Publishing, 2000.
- [38] C. Anderson, *The Long Tail*. London: Random House Books, 2006.
- [39] M. Gladwell, *The Tipping Point*. London: Abacus, 2000.
- [40] R. E. Caves, *Creative Industries*. Cambridge: Harvard Univ. Press, 2000.
- [41] P. Bourdieu, *The Rules of Art*. Stanford: Stanford Univ. Press, 1996.
- [42] C. Wiertz and K. de Ruyter, *Organization Studies*, vol. 28, no. 3, pp. 347-376, 2007.
- [43] M. Asensio and V. Hodgson, in *Communities of Practice: Trends in Communication Series*, vol. 8., M. Huysman and P. van Baalen, Eds. Amsterdam: Boom, 2001, pp. 65-77.
- [44] P. Bonacich, *Am. J. Soc.*, vol. 92, no. 5, pp. 1170-1182, 1987.
- [45] B.A. Huberman, D.M. Romero, and F. Wu. *First Monday*, vol. 14, no.1, 2009.
- [46] M. E. J. Newman. *Eur. Phys. J. B*, vol. 38, pp. 321-330, 2004.
- [47] L. Danon, J. Duch, A. Diaz-Guilera and A. Arenas, *J. Stat. Mech.*, P09008, 2005.
- [48] E. R. Dugundji and J. L. Walker, *Transportation Research Record*, no. 1921, Washington, DC, 2005.