Catching Social Butterflies: Identifying Influential Users of an Event-Based Social Networking Service

Jonathan Popa, Kusha Nezafati, Yulia R. Gel, John Zweck
Department of Mathematical Sciences
The University of Texas at Dallas
Richardson, TX, USA
{jonathan.popa1, Kusha.Nezafati, ygl, zweck}@utdallas.edu

Georgiy Bobashev
RTI International
Research Triangle Park, NC, USA
bobashev@rti.org

***Abstract*—Online social media information is often used as a proxy for unavailable or partially observed data on networks of offline contacts. This, in turn, requires an understanding of how close the proxy online structure is to the "true" offline social network. Social media tools such as Meetup that collect information about *both* online networks and their offline counterparts are of particularly importance as they shed more light on the (dis)similarity of online and offline contacts and highlight its potential causes. In this paper we examine structural (dis)similarities of the Meetup online and offline data, with a particular focus on geographical differences. We introduce a new measure called the *event score* to assess connections made by the most socially active individuals, or *social butterflies*. We apply the new social activity metric to determine which sorts of events are attended most by social butterflies and to evaluate how this aspect of the network structure differs across US cities.**

***Index Terms*—big data; event-based social networks**

I. Introduction

It is well known that the manner in which an agent such as news, virus, behavior, etc, spreads throughout a community strongly depends on the structure of connections between individuals in the social system [1, 11, 13, 35]. This phenomenon is largely due to the fact that individuals tend to adopt new ideas or behavior that are similar to the choices of their peers [22, 24]. Nowadays modeling diffusion on interaction networks is therefore attracting ever increasing attention in the social sciences, statistics and computer science, especially in view of its key role in developing efficient strategies for viral marketing and for risk mitigation against emerging and re-emerging diseases with a high virulence such as Ebola, Zika, swine and avian flu. One of the primary and fundamental questions in this field is to reliably evaluate and model the underlying structure of social communication networks, that is, *to understand how people interact with each other*. While there are numerous studies on structures of online social communication networks, because of limited data sources, analysis of offline interaction structures is largely restricted to closed or compact populations such as prisons, schools and hospitals [5, 14, 16, 20, 30].

One of the largest and most accurate studies reflecting a general structure of social mixing patterns is the Polymod survey of 7,290 individuals in eight European countries [28]. However, in view of cultural, city infrastructure, socio-demographic and other differences, it is unclear how transferrable the Polymod mixing patterns are to other geographical areas. Another way to collect data on social communication is via various wearable wireless devices, for instance, mobile phones, that measure the physical proximity between individuals [2, 17, 34, 39]. On the other hand, there is a growing trend in using data from online social media as a "proxy" for unobserved offline contacts [2, 3, 10, 15, 23, 26, 33, 38]. The wide availability and popularity of online platforms make such a surrogate approach an attractive data collection channel in a broad range of applications, especially, if no offline data are available [6]. This, in turn, requires the understanding of how representative or close the "proxy" online interaction structure is to that of the "true" offline communication network. Hence, social media tools that gather data about *both* online networks and their offline counterparts are of particular importance as they shed light on the (dis)similarity of online and offline contacts and highlight its potential causes.

In this regard, publicly available data from the social networking service *Meetup* are *arguably* unique for these purposes in North America, as Meetup offers information on both the online and offline interactions of its members. The online network is constructed by connecting Meetup users who belong to the same interest group, while the offline network is constructed by connecting Meetup users who attend the same event. Liu et al. [26] perform an analysis of the Meetup dataset to examine properties of event-based social networks. Their study suggests that the online network tends to be significantly denser than the offline network, and the degree distributions of both the online and offline networks are more heavy tailed than a classical power law distribution. The findings in [26] also indicate that Meetup users are likely to participate primarily in events within 10 miles of their registered location, leading to clustering. Most recently, the Meetup data have been analyzed in a context of a social event recommendation system but without an analysis of offline vs. online network structures [7].

This paper constitutes a pilot study that further explores structural (dis)similarities of the Meetup online and offline data, with a particular focus on geographical differences. Rather than examining the entire US network, we refine our scope to the level of six cities in the United States: Chicago, Dallas, Los Angeles, Miami, New York City, and

Philadelphia. To our knowledge, analysis of the Meetup data at a city-level scale has never been performed before. Our paper is motivated by two overarching questions. First, we are interested to evaluate whether online and offline networks at a city level exhibit similar patterns. Our study concludes that the degree distributions of the online networks are more dispersed than the offline networks across all cities, but that all the distributions share a heavier tail than the power law distribution. We also conclude that a user in any of the six cities we considered is more likely to join many groups online than to interact with several groups offline.

Second, it is well known that extremely socially active individuals can have a large impact on the spread of an agent throughout a network [19, 25, 26]. Therefore, we are interested to determine whether socially active individuals behave differently in different cities. For many classes of networks, the influence of socially active individuals is encoded in the tails of the degree distribution. However, the online and offline Meetup networks are dominated by clique-based structures since, absent any more fine-grained data, the online networks are defined to include connections between all members of a given online group and the offline networks include connections between all members that attend the same event. The question is therefore to understand the web of connections between the different groups in the online network and the different events in the offline network. We suggest that this web of connections is strongly influenced by *social butterflies*, i.e. by the most socially active individuals. We introduce a new measure to evaluate connections made by social butterflies, namely the notion of an *event score*, and apply it to determine which sorts of events are attended most by social butterflies. Our results show that a large proportion of the attendees at smaller events are social butterflies, whereas there tends to be a much smaller proportion of social butterflies at larger events. (Put another way, people who want to socialize and meet someone to date are more likely to succeed at a smaller event.) We also find that the distribution of such event scores depends on the choice of city and persists even if we account for city population and population density. While we can offer no clear explanation for this result, our findings suggest that there may be city-specific factors that influence how agents propagate through a social network.

Despite being a pilot study, these findings are thus of particular interest for city-specific biosurveillance and early awareness platforms for real-time tracking and forecasting of infectious diseases, based on harnessing various traditional data sources (i.e., data from public health agencies) and non-traditional information (i.e., health-related online social media such as Twitter, Google trends, etc) [see, e.g., 8, 12, 21, 27, 29, 31, and references therein]. The differences our study highlights between online and offline social networks across cities also provides valuable insights for the development of more effective targeted marketing strategies, survey sampling procedures, disaster warning tools, and even dating and matchmaking agencies [4, 9, 18, 32].

The paper is organized as follows. In Section 2 we provide a description of the Meetup data, and in Section 3 we discuss construction of online and offline networks and evaluate the respective degree distributions. In Section 4 we introduce a new notion of social activity, that is, an event score (ES) and robust event score (RES), and in Section 5 we study differences in event scores among the six US cities. The paper concludes with discussion and future work (Section 6).

II. Description of Meetup Data

Meetup is a social networking service that allows users to form online groups and records users' RSVPs to offline events. Together these functionalities allow for the generation of datasets containing information on both online interactions in groups and offline interactions at events. For this study, we used a publicly available data set compiled by Liu et al. [26] from meetup.com. This data contains information on the group membership and event attendance of over four million users, collected between October 2011 and January 2012. To ensure anonymity, each user is assigned a unique user ID. Similarly, each group is assigned a unique group ID and each event is assigned a unique event ID. The files in the dataset can be partitioned into three categories: *geographic*, *network*, and *tag*.

The two files in the geographic category give user locations and event locations arranged as a matrix with three columns. Each row contains a user ID (resp. event ID) followed by the longitude and latitude coordinates of the user's registered home location (resp. event's location).

There are three files in the network category, each arranged as a matrix with two columns. The first file relates users and groups, with each row consisting of a user ID followed by a group ID. This file defines the *members* of each group, that is, we say a user is a member of a group if the associated user and group IDs appear on the same row of this matrix. The second file relates users and events, each row containing a user ID and event ID. This file defines the *attendees* at each event, that is, we say a user attends an event if the associated user and event IDs appear on the same row of this matrix. We observe that the set of attendees at an event is the set of users who RSVP'd "yes" to the event, and as such only gives an indication of who physically attended the event. Some users who RSVP'd "yes" may not attend the event, while other users who do not RSVP "yes" may nevertheless still physically attend the event. The third network file relates events and groups, each row containing an event ID and a group ID. This file gives information on which events were hosted by which groups. Each event can be hosted by at most one group and some events have no hosting group. Furthermore, users are able to attend events hosted by groups of which they are not members.

There are three data files in the tag category, each arranged as a matrix. A *tag* is a descriptive identifier which can be assigned to a user or group. For example, a group of baseball fans in New York may be tagged "baseball" or "Yankees". Each row of the first file associates a number (the tag ID) to each descriptive tag (eg. "baseball"). Each row of the second (resp. third) file consists of a user ID (resp. group ID) followed

by a tag ID. This allows users and groups to be identified by their interests.

III. Online vs. Offline Networks

Construction of Networks We used the Meetup data set to study the network structure for six US cities: Dallas, New York, Los Angeles, Miami, Chicago, and Philadelphia. These cities were selected because of their large and diverse populations. For each city we compiled a list of *residents* who are Meetup users whose registered location was within the city. For simplicity, the geographical extent of each city was defined to be the circle in longitude-latitude space that is centered at the city's geographic center and whose area is equal to the area of the city. The geographical centers and areas of each city were obtained from wikipedia.

For each city we constructed two networks from the Meetup data, an online network and an offline network. In each network a *vertex* is a resident and an *edge* represents contact between two residents. For the online network of a city, we construct a single edge between two vertices if there is at least one group to which the corresponding residents both belong. For the offline network of a city, we construct a single edge between two vertices if there is at least one event that the corresponding residents both attend. Our assumption of complete mixing at each event is reasonable since the average event attendance ranged between 3.5 and 4.3 for all six cities, which suggests that most events were small in size.

An inherent characteristic of these networks is that the set of residents in a group or at an event forms a clique. A clique is a group of vertices that are fully-connected to each other. Figure 1 is a toy example of four cliques connected by common vertices. These cliques would represent events in an offline network or groups in an online network. Table III presents the summary statistics on the online and offline Meetup data for the six considered cities.

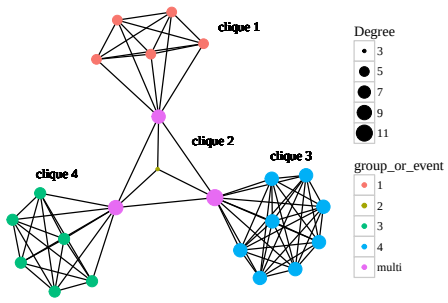

Fig. 1. Toy example of network structure

Evaluating Degree Distributions In Figure 2, we show log-log plots of the degree distributions of the online networks in five of the cities we studied: Dallas, New York, Los Angeles, Miami, Chicago, and Philadelphia. (Due to computational limitations we do not present a network distribution for New York City.) The large deviations from the linear fits (solid blue lines) suggest that these degree distributions are more heavy tailed than the classical power-law distribution, which mirrors patterns over the entire United States [26]. The city whose online degree distribution is closest to a power law distribution is Miami, which is the smallest of the five cities.

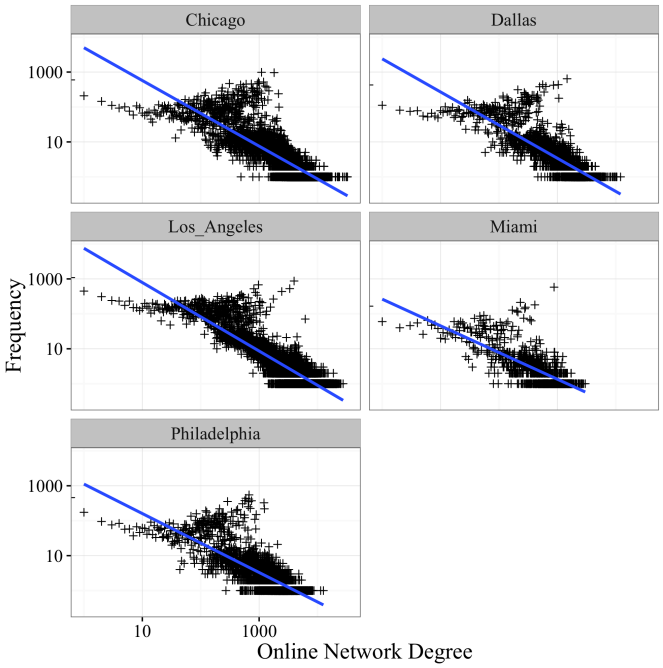

Fig. 2. Online degree distribution for major cities

Figure 3 shows the respective log-log plots of the degree distributions of the offline networks. The results suggest that similarly to the online networks, the offline degree distributions do not follow the usual power law distribution either but are noticeably more heavy tailed.

IV. Quantifying Social Activity

Our analysis of social activity of Meetup users and their interaction patterns are based on the two metrics: *active membership* and *event score*.

A user is an *active* member in a group if the user attends an event hosted by the group. Since events are not restricted to members of the hosting group, a user can be active in a group without being an online member of the group. We also found that some events had no hosting group. We treated these events as all being hosted by a single online proxy group. In Figure 4 we compare the number of memberships and active memberships of the residents in each city. Each data point in the plot represents a resident. The number of memberships for each user was determined by scanning through the user/group ID matrix, and counting the number of times a user's ID appeared. To calculate the number of active memberships of a user, we identified all events the given user attended and recorded the ID of the groups hosting these events. The number of active memberships is then the cardinality of this set of group IDs.

The purpose of the active membership metric is to compare how users behave online versus offline. We observe that a

TABLE I
SUMMARY STATISTICS OF ONLINE AND OFFLINE MEETUP DATA IN THE SIX US CITIES.

City	Miami	Dallas	Philadelphia	Chicago	LA	NYC
# of users	7,537	29,218	26,979	66,148	78,884	314,510
# of events	14,505	37,521	33,339	68,650	107,468	227,934
# of groups	249	623	565	1,126	2,093	4,753

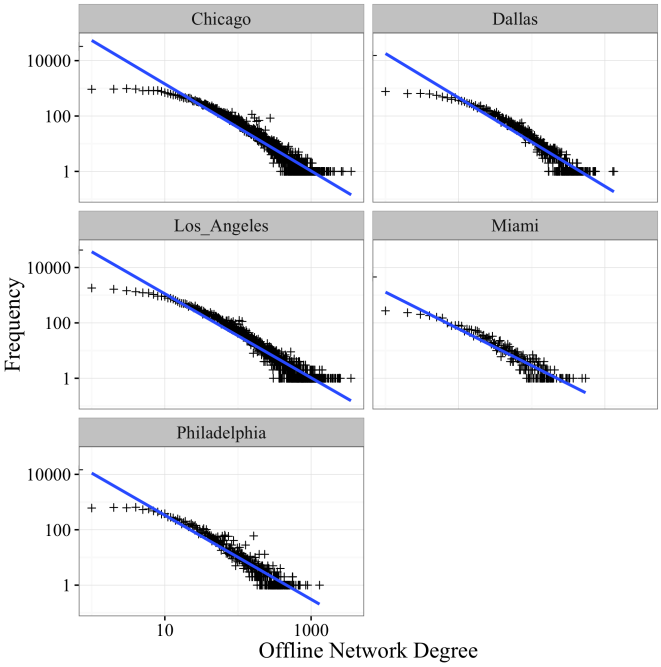

Fig. 3. Offline degree distribution for major cities

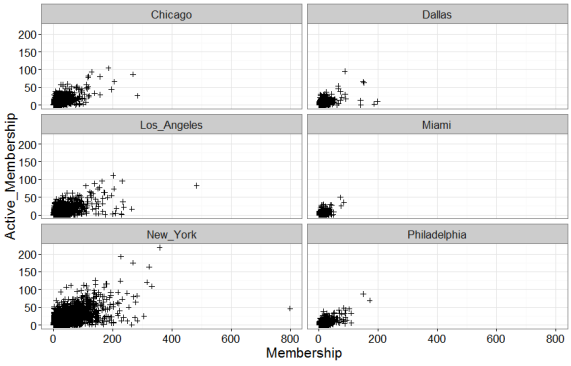

Fig. 4. Membership vs Active Membership

number of users who hold membership in many groups is larger than a number of users who have many active group memberships. This result is to be expected since it is easier to join an online group compared to the effort of traveling to attend offline events.

To quantify offline social activity, we define the *measure of social activity* of a user to be the number of events the user attended. We refer to those users with larger measures of social activity as *social butterflies*. It is important to note that since there exists no widely agreed upon measure of social activity, such measures often need to be tailored to the structure of dataset. Our definition is ambiguous in the sense that a social butterfly who attends N events with the same group of people (*e.g.*, weekly meetings of a chess club) has the same measure of social activity as a user who attends N different sorts of events hosted by different groups. However, we argue that despite this ambiguity, this metric of social activity is plausible since in both cases such social butterflies have a higher probability of transmitting an agent than users with lower measures of social activity. For example, a social butterfly who attends many chess club events tends to transmit information from one club meeting to the next, while a social butterfly who attends meetings hosted by many different online groups tends to spread information from one group of users to another. In the future we plan to distinguish social butterflies with respect to the size and type of events they tend to participate in.

Our notion of social butterfly is also closely related to the notion of the betweenness centrality of a node in a social network. The betweenness of a node counts the number of shortest paths between pairs of nodes that pass through the given node, relative to the total number of shortest paths between each pair [36]. Individuals with larger betweenness centrality play an important role in the transmission and diffusion of ideas and infectious agents. Another closely related notion is that of an influential user of a social network that can be measured as the impact on others' activity levels (see discussion in [36, 37] and references therein), which requires data on how information is propagated in the network as a function of time. Since the events in the *Meetup* database we relied on for this study include neither the time at which the event occurred nor a more detailed graph structure, it is not feasible to calculate the influence or the conventional betweenness centrality of *Meetup* users. For these reasons, and because of the particular structure of the Meetup offline network described in Section 3, for the results in this paper we use our measure of social activity as a proxy for betweenness centrality and influence.

In Table II we show the percentiles of user attendance. We observe that most users have only attended one or two events, while the most active users, the social butterflies, attended hundreds or thousands of events. The percentiles of event sizes, that we show in Table III, reveal that most events are attended by at most four users, and the largest events are attended by hundreds of users.

To examine whether most socially active individuals tend to cluster together, that is, whether social butterflies are more

TABLE II
USER ATTENDANCE (AGGREGATED OVER ALL SIX CITIES).

Percentiles	0%	25%	50%	75%	90%	99%	100%
User Attendance	1	1	2	6	16	95	4829

TABLE III
EVENT SIZE (AGGREGATED OVER ALL SIX CITIES).

Percentiles	0%	25%	50%	75%	90%	99%	100%
Event Size	2	2	4	8	15	56	722

likely to attend larger events or smaller events, we introduce two new measures of event activity: event score and robust event score. In particular, we assign such a score to each event to evaluate social activity of its attendees. Given an event, we define the *event score* (ES) as

$$\text{ES} = \frac{1}{N}\sum_{j=1}^{N} X_j, \tag{1}$$

where N is the number of attendees at the given event and X_j is the measure of social activity of the j-th attendee of the given event.

In Figure 5 we show event score vs. event attendance for the six cities in our study. Rather than showing a scatter plot of event score—event attendance pairs, we plot the 95-th percentile of the event score as a function of event attendance. Specifically, these curves are obtained by applying an additive quantile regression model for the 95-th percentile. The results indicate that events with smaller attendance tend to have higher event scores, whereas larger events have noticeably lower event scores. Since the event score is an average of the attendees' social activity measure, a high event score indicates that an event is attended by relatively large number of social butterflies. Consequently, we conclude that social butterflies tend to be attracted to a larger number of smaller events and to be surrounded by similarly active individuals; and in reverse that larger events tend to be dominated by less social individuals. This is further supported by our second robust median-based scoring method (see discussion below). We observe that this conclusion needs to be balanced with the data in Table III which shows that larger event sizes are more frequent than smaller ones. However, events are ultimately organized by people, and the high frequency of smaller meetings may result from a tendency of social butterflies to organize smaller rather than larger events. This finding has significant implications for social diffusion models, and in particular to understanding spread of agents through a societal structure.

The second scoring method, the *robust event score* (RES) is given by

$$\text{RES} = \text{med}\{X_j : j = 1, 2, ..., N\} \tag{2}$$

The idea of RES to diminish the impact of outliers when evaluating social activity. In Figure 6 we show the 95-th percentile of the RES as a function of the event attendance. We use the same cut-offs as in Figure 5. We observe a similar

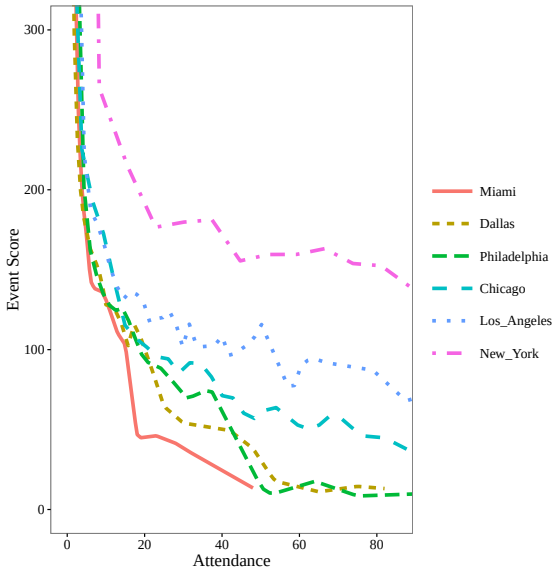

Fig. 5. Event Score vs Attendance

relationship between RES and event attendance as we find for ES and event attendance. The results shown in Figures 5 and 6 suggest that there are noticeable differences between cities. To further assess these differences in social activity, in Figures 7 and 8 we show boxplots of ES and RES, respectively, for each city where the cities are ordered in terms of their population size. We find that despite being the city with the least population, Miami exhibits almost the same variability as the second largest city, Los Angeles. In turn, Dallas (the second smallest city in terms of population size) has the most concentrated distribution, while New York City (the largest city) exhibits the most dispersed distribution. In the next section we present the results of a formal hypothesis test for the homogeneity of ES and RES.

V. STATISTICAL ANALYSIS OF EVENT SCORES ACROSS U.S. CITIES

Here we formally test the hypothesis that Meetup users in the six major US cities behave similarly (H_0) vs. that there exist differences in their social dynamics. We start from evaluating mean ES and RES among the cities. Since Dallas, New York, Los Angeles, Miami, Chicago, and Philadelphia

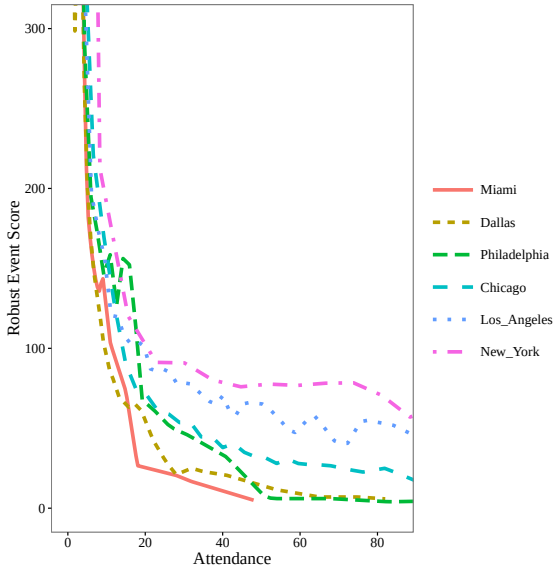

Fig. 6. Robust Event Score vs Attendance

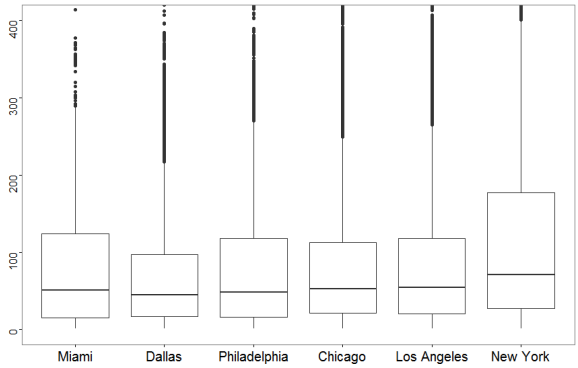

Fig. 7. Event Score Box Plot

have different population sizes and densities, we start by accounting for these demographic factors as covariates.

In particular, we first regress event scores vs. population size and density, which as expected, suggest significance of both demographic variables. We then apply a one-way analysis of variance (ANOVA) to the resulting residuals to test for difference among cities. The ANOVA test rejects the null hypothesis, and the city effect is found to be highly statistically significant. To account for nonnormality and heteroscedasity of residuals, we also apply logarithmic transformations prior to ANOVA as well as a non-parametric distribution-free Kruskal-Wallis test. In both cases, we find that the city effect remains highly statistically significant. In Figure 9 we show boxplots of the model residuals subject to the logarithm transformation of ES and accounting for demographic factors. We find that Miami (the city with the lowest population) exhibits almost the same variability as Philadelphia (the third smallest city). At the same time, the main body of residuals distributions (i.e., the area between the lowest and highest quartiles) in Dallas, Chicago, Los Angeles, and New York is similar. Analogous

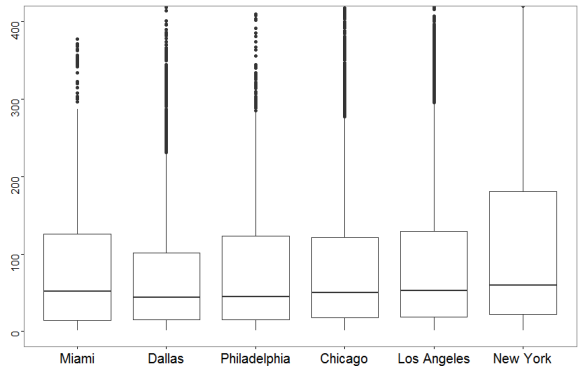

Fig. 8. Robust Event Score Box Plot

results are obtained for RES and are omitted for brevity.

At this point, we cannot offer a plausible explanation for such dissimilarities among Meetup users who reside in different cities. We may hypothesize that such dissimilarities might be due to city infrastructure, ethnic and other socio-demographic differences etc. (To further investigate these patterns, we are currently extending our study to other large metropolitan areas.) Nevertheless, these findings indicate that propagation of news, viruses, behaviors and trends are likely to differ substantially between cities and need to be accounted for in social diffusion models and survey designs.

Fig. 9. Residuals Box Plot

VI. Conclusion

The availability of the Meetup data opens new horizons in understanding the structure and organization of both online and offline social networks. The current pilot project is an attempt to shed some new light on the (dis)similarity of patterns of online and offline interactions in various US cities. Despite being a preliminary analysis, the results we obtained have a number of important implications. First, for most social networks obtaining the complete network is problematic and the assessment is usually based on a network sample. As shown in [4], sampling methodology strongly affects the bias of the network estimates drawn from the sample. The effect of such bias could be practically studied when the entire network is known. Understanding the structure of close to

complete online and offline Meetup networks provides a way to compare, evaluate and control sampling methods when *only a sample* of the networks is known. One of the examples is sampling based on the Meetup events. If the majority of events are small in size, sampling from a small group will likely lead to a biased representation due to dominance by "social butterflies". Second, differences in social activity among cities need to be accounted in constructing offline networks of social contacts based on synthetic populations [8, 27, 29, 31].

In the future we plan to extend this analysis in multiple directions. To increase understanding of the relationships between the structure of online and offline Meetup datasets, we will evaluate event scores with respect to the sizes and types of events and interest groups, and will study space-time and multi-attribute networks of Meetup events and users.

By using quantile regression of offline vs. online degree distributions, we can obtain city-wise factors for relating offline structure with its proxy. Another interesting direction is to investigate further stratification of the Meetup data in terms of group interests and its respective dynamics across the United States. Finally, we plan to use graph matching procedures to relate other online sources such as Facebook and Twitter to that of online Meetup and then to explore utility of online Meetup as a link between a richer set of online networks and the respective offline interaction patterns.

ACKNOWLEDGMENT

We would like to thank the anonymous reviewers for their detailed and constructive comments and suggestions that helped to improve the manuscript. This project was supported under the NSF Enriched Doctoral Training Program entitled "Team Training Mathematical Scientists Through Industrial Collaborations", DMS Award # 1514808.

REFERENCES

[1] M. A. Andrews and C. T. Bauch. The impacts of simultaneous disease intervention decisions on epidemic outcomes. *Journal of theoretical biology*, 395:1–10, 2016.

[2] M. Atzmueller and K. Hilgenberg. Towards capturing social interactions with sdcf: An extensible framework for mobile sensing and ubiquitous data collection. In *Proceedings of the 4th International Workshop on Modeling Social Media: Mining, Modeling and Recommending 'Things' in Social Media, MSM 2013*, 2013.

[3] L. Backstrom and J. Kleinberg. Romantic partnerships and the dispersion of social ties: A network analysis of relationship status on facebook. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW*, pages 831–841, 2014.

[4] G. Bobashev, R. J. Morris, and M. Goedecke. Sampling for global epidemic models and the topology of an international airport network. *PLoS ONE*, 3(9), 2008.

[5] F. Bonchi, C. Castillo, A. Gionis, and A. Jaimes. Social network analysis and mining for business applications. *ACM Transactions on Intelligent Systems and Technology*, 2(3), 2011.

[6] M. W. Carroll, D. A. Matthews, J. A. Hiscox, and et al. Temporal and spatial analysis of the 2014-2015 ebola virus outbreak in west africa. *Nature*, 524(7563):97–101, 2015.

[7] C. C. Chen and Y. . Sun. Exploring acquaintances of social network site users for effective social event recommendations. *Information Processing Letters*, 116(3):227–236, 2016.

[8] J. Chretien, D. George, J. Shaman, R. A. Chitale, and F. E. McKenzie. Influenza forecasting in human populations: A scoping review. *PLoS ONE*, 9(4), 2014.

[9] C. Dijkmans, P. Kerkhof, and C. J. Beukeboom. A stage to engage: Social media use and corporate reputation. *Tourism Management*, 47:58–67, 2015.

[10] R. Du, Z. Yu, T. Mei, Z. Wang, Z. Wang, and B. Guo. Predicting activity attendance in event-based social networks: Content, context and social influence. In *UbiComp 2014 - Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 425–434, 2014.

[11] W. Duan, Z. Fan, P. Zhang, G. Guo, and X. Qiu. Mathematical and computational approaches to epidemic modeling: a comprehensive review. *Frontiers of Computer Science*, 9(5):806–826, 2015.

[12] A.F. Dugas, M. Jalalpour, Y. Gel, S. Levin, F. Torcaso, T. Igusa, and R.E. Rothman. Influenza Forecasting with Google Flu Trends. *PLOS One*, 8:e56176, 2013.

[13] E. Estrada, F. Kalala-Mutombo, and A. Valverde-Colmeiro. Epidemic spreading in networks with nonrandom long-range interactions. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 84(3), 2011.

[14] L. M. Glass and R. J. Glass. Social contact networks for the spread of pandemic influenza in children and teenagers. *BMC Public Health*, 8, 2008.

[15] P. A. Grabowicz, J. J. Ramasco, E. Moro, J. M. Pujol, and V. M. Eguiluz. Social features of online networks: The strength of intermediary ties in online social media. *PLoS ONE*, 7(1), 2012.

[16] C. G. Grijalva, N. Goeyvaerts, H. Verastegui, K. M. Edwards, A. I. Gil, C. F. Lanata, and N. Hens. A household-based study of contact networks relevant for the spread of infectious diseases in the highlands of peru. *PLoS ONE*, 10(3), 2015.

[17] B. Guo, Z. Wang, Z. Yu, Y. Wang, N. Y. Yen, R. Huang, and X. Zhou. Mobile crowd sensing and computing: The review of an emerging human-powered sensing paradigm. *ACM Computing Surveys*, 48(1), 2015.

[18] M. Gnois, C. L. Vestergaard, C. Cattuto, and A. Barrat. Compensating for population sampling in simulations of epidemic spread on temporal contact networks. *Nature Communications*, 6, 2015.

[19] S. Hajian, T. Tassa, and F. Bonchi. Individual privacy in social influence networks. *Social Network Analysis and*

Mining, 6(1):1–14, 2016.

[20] M. S. Handcock and K. J. Gile. Modeling social networks from sampled data. *Annals of Applied Statistics*, 6(1):5–25, 2012.

[21] K. S. Hickmann, G. Fairchild, R. Priedhorsky, N. Generous, J. M. Hyman, A. Deshpande, and S. Y. Del Valle. Forecasting the 2013?2014 influenza season using wikipedia. *PLoS Computational Biology*, 11(5), 2015.

[22] M. Jackson and L. Yariv. Diffusion on social networks. *Économie Publique*, 16:3–16, 2005.

[23] R. Jurdak, K. Zhao, J. Liu, M. AbouJaoude, M. Cameron, and D. Newth. Understanding human mobility from twitter. *PLoS ONE*, 10(7), 2015.

[24] D. Kempe, J. Kleinberg, and . Tardos. Influential nodes in a diffusion model for social networks. In *Lecture Notes in Computer Science*, volume 3580, pages 1127–1138, 2005.

[25] A. Liccardo and A. Fierro. Multiple lattice model for influenza spreading. *PLoS ONE*, 10(10), 2015.

[26] X. Liu, Q. He, Y. Tian, W. Lee, J. McPherson, and J. Han. Event-based social networks: Linking the online and offline social worlds. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1032–1040, 2012.

[27] L. J. Martin, B. E. Lee, and Y. Yasui. Google flu trends in canada: A comparison of digital disease surveillance data with physician consultations and respiratory virus surveillance data, 2010-2014. *Epidemiology and infection*, 144(2):325–332, 2016.

[28] J. Mossong, N. Hens, M. Jit, P. Beutels, K. Auranen, R. Mikolajczyk, M. Massari, S. Salmaso, G. S. Tomba, J. Wallinga, J. Heijne, M. Sadkowska-Todys, M. Rosinska, and W. J. Edmunds. Social contacts and mixing patterns relevant to the spread of infectious diseases. *PLoS Medicine*, 5(3):0381–0391, 2008.

[29] EO. Nsoesie, M. Marathe, and JS. Brownstein. Forecasting peaks of seasonal influenza epidemics. *PLOS Currents Outbreaks*, 2013.

[30] G. E. Potter, M. S. Handcock, I. M. Longini, and M. Elizabeth Halloran. Estimating within-household contact networks from egocentric data. *Annals of Applied Statistics*, 5(3):1816–1838, 2011.

[31] L. L. Ramrez-Ramrez, Y. R. Gel, M. Thompson, E. de Villa, and M. McPherson. A new surveillance and spatio-temporal visualization tool SIMID: SIMulation of infectious diseases using random networks and GIS. *Computer methods and programs in biomedicine*, 110(3):455–470, 2013.

[32] V. Sánchez, N. Muñoz Fernández, and R. Ortega-Ruíz. "cyberdating q-a": An instrument to assess the quality of adolescent dating relationships in social networks. *Computers in Human Behavior*, 48:78–86, 2015.

[33] S. Scellato and C. Mascolo. Measuring user activity on an online location-based social network. In *2011 IEEE Conference on Computer Communications Workshops, INFOCOM WKSHPS 2011*, pages 918–923, 2011.

[34] J. Stehl, N. Voirin, A. Barrat, C. Cattuto, L. Isella, J. . Pinton, M. Quaggiotto, W. van den Broeck, C. Régis, B. Lina, and P. Vanhems. High-resolution measurements of face-to-face contact patterns in a primary school. *PLoS ONE*, 6(8), 2011.

[35] T. Takaguchi and R. Lambiotte. Sufficient conditions of endemic threshold on metapopulation networks. *Journal of theoretical biology*, 380:134–143, 2015.

[36] L. Tang and H. Liu. Community detection and mining in social media. *Synthesis Lectures on Data Mining and Knowledge Discovery*, 2(1):1–137, 2010.

[37] M. Trusov, A. V. Bodapati, and R.E. Bucklin. Determining influential users in internet social networks. *Journal of Marketing Research*, XLVII:643–658, 2010.

[38] C. Wilson, B. Boe, A. Sala, K. P. N. Puttaswamy, and B. Y. Zhao. User interactions in social networks and their implications. In *Proceedings of the 4th ACM European Conference on Computer Systems, EuroSys'09*, pages 205–218, 2009.

[39] D. Yu, R. C. Blocker, M. Y. Sir, M. S. Hallbeck, T. R. Hellmich, T. Cohen, D. M. Nestler, and K. S. Pasupathy. Intelligent emergency department: Validation of sociometers to study workload. *Journal of medical systems*, 40(3):1–12, 2016.