2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)

Social Event Network Analysis: Structure, Preferences, and Reality

Martin Atzmueller, Tom Hanika, and Gerd Stumme
Research Center for Information System Design
Chair of Knowledge and Data Engineering
University of Kassel, Germany
{atzmueller,hanika,stumme}@cs.uni-kassel.de

Richard Schaller
AG Digital Humanities, Computer Science, University of Erlangen-Nuremberg, Germany
richard.schaller@fau.de

Bernd Ludwig
I:IMSK
University of Regensburg, Germany
bernd.ludwig@ur.de

***Abstract*—This paper focuses on the analysis of socio-spatial data, i. e., user–performance relations at a distributed event. We consider the data as a bimodal network (i. e., model it as a bipartite graph), and investigate its structural characteristics towards a social network. We focus on plans of the participants (expressed by preferences) and their fulfilment, and propose measures for matching preference and reality. We specifically analyse behavioural patterns w.r.t. distinct user and performance groups. We utilise real-world data collected at the Lange Nacht der Musik (Long Night of Music) 2013 in Munich.**

1. Introduction

In cities, distributed cultural events are becoming more and more widespread - enabling the selection and involvement in a large variety of performances for the participants. The analysis of the respective relations between participants and particular parts of an event (performances) can yield important insights into the event structure, as well as provide a detailed view on user interests (expressed by preferences) and individual/collective behavioural attendance patterns.

Objectives. Both perspectives, i. e., structure and behaviour, are the two main research foci addressed in this paper. We analyse data collected from a distributed event, the Lange Nacht der Musik, Long Night of Music - a cultural event organised in the city of Munich, Germany, which we model using methods from social network analysis. We model user–performance relations corresponding to the *planned behaviour* (preference) and the *real behaviour* as bipartite graphs. Then, we analyse these concerning their structural characteristics. In addition, we analyse behavioural patterns with respect to the fulfilment of the planned behaviour using formal concept analysis [1] and subgroup discovery [2].

Contribution. (1) We focus on the social structure of the distributed event where preferences and visiting behaviour can be regarded as expressing socio-spatial characteristics of the users. We analyse the bimodal network of user–performance relations, and show that this social event network conforms to the characteristics of a social network. (2) In addition, we demonstrate that we can identify *distinct* behavioural patterns w.r.t. groups of users and performances. (3) In doing that, we furthermore propose and demonstrate a set of methods for the analysis of such network models.

The rest of the paper is structured as follows: Section 2 discusses related work. Then, Section 3 describes our data. Sections 4-5 tackle the structural and behavioural analysis tasks, respectively. Finally, Section 6 concludes with a discussion and outlines interesting directions for future work.

2. Related Work

Event-based networks are captured by interactions between people, either online [3] or offline, e. g., [4], [5]. A general view on modelling and mining of ubiquitous and social data is given in [6] focusing on social interaction networks captured during certain events, e. g., during conferences. Offline contact patterns, and their underlying mechanisms are analysed, e. g., relating to roles and interactions between communities [7], or their dynamics [4]. In contrast to those approaches, we do not focus on explicit offline event-based networks, nor on (location-based) online social networks. We consider *implicit* location-based networks that are formed by participants of a distributed cultural event that visit different performances. We show that different induced networks, e. g., links between users or performances, contain important socio-spatial characteristics. A similar approach has been conducted in [8], however concerning the semantics of implicit *online* user interactions, in contrast to our offline setting. In addition, we do not consider the semantics of the (implicit) interactions, but focus on structural aspects as well as behavioural patterns.

In [9], [10], the authors investigated how the information search behaviour of individual users utilizing an electronic guide influences the behaviour while planning the attendance of a distributed event and the behaviour while visiting it. Furthermore, the authors provide evidence that during the event the information search behaviour induces modifications of the attendance behaviour with respect to the planning behaviour. Here, we extend this kind of behavioural analysis of individual users to behavioural patterns: We first build a social network from log data of users observed during a distributed event. Then, we reconstruct typical behaviour of user groups from properties of that social network.

Furthermore, we do not only focus on the analysis: Instead, we present a methodology for the analysis of such social (distributed) event networks, e. g., concerning structure, social characteristics, and distinct behavioural patterns.

IEEE/ACM ASONAM 2016, August 18-21, 2016, San Francisco, CA, USA
978-1-5090-2846-7/16/$31.00 ©2016 IEEE
PREPRINT

3. Description of Data and Modelling

The Lange Nacht der Musik (Long Night of Music; LNM) is an annual cultural event organised in the city of Munich. In addition to a diverse range of pubs, discotheques and clubs, other cultural venues, such as churches and museums open their doors for one evening in May to host musical performances. On May 11th 2013, approximately 20,000 people visited a total of 212 available performances at 113 distinct locations, that were dispersed across the city.

LNM App. Since each individual visitor can only visit a small fraction of these performances, careful planning is necessary. For this task, an Android app was developed by the University of Erlangen-Nuremberg that assists in finding performances of interest and creates itineraries that minimise travel time while maximising time spent at a performance. The app provides personalised recommendations to the user [11], allows faceted browsing of the performances and finally creates a personalised plan for visiting a set of performances based on the user's preferences. At any moment the user can decide to edit the remainder of the plan, e. g., by removing a performance, inserting an performance at a specific position/at its best position, rearranging the order of performance visits or by changing a performance's visit duration. The app was offered in the Google Play Store.

Data. Overall 1159 visitors used the app, shaping the set of all user IDs U. For those we logged all user interactions. Of all users, 612 actually rated performances from the performance set H for possible tour inclusion. They did so by assigning a preference value from the *preference set* $P = \{-1, 0, 1, 2\}$, where -1 = "no rating", 0 = "Don't want to visit", 1 = "Would like to visit, if it fits into the tour", and 2 = "Want to visit in any case", meant. This ascending encoding will enable us to calculate easily with the preferences later on, especially to omit "no rating". We further call this data set the *preference data set (DB_P)*. Additionally, we logged GPS (Global Positioning System) data in order to track users' actual visits. As not all users had their GPS enabled, we were only able to reconstruct the visits of 111 out of the 1159 visitors. We call this the *attendance data set (DB_A)*.

Modelling. From the data described above we inherit multiple graph structures. In particular we model DB_P as an edge-weighted bipartite graph with $U \cup H$ as the set of vertices and some edge set $E \subseteq U \times H$. The weight function will be $w\colon U \times H \to P$, which maps a user u and a performance h to the preference value $p \in P$, which was assigned by u to h prior to the LNM. Using this we define the edge set $E \coloneqq \{(u,h) \in U \times H \mid w(u,h) \geq 0\}$, i. e., there is an edge between u and h if u assigned a preference to h. Obviously, taking users without any preference value into account would lead to a vast amount of singleton components. They are therefore omitted, as well as performances not preferred by any user. We model this by the restricted subsets $U_P \subseteq U$ and $H_P \subseteq H$. This results in the *preference graph* $\mathcal{P} \coloneqq (U_P \cup H_P, E_p, w)$ where $E_P \subseteq E$ is the induced subset on $U_P \cup H_P$. This graph is connected and has 824 vertices (612 users and 212 performances).

By focusing on the set of positive preferences, i. e., $E_D \coloneqq \{(u,h) \in U \times H \mid w(u,h) \geq 1\}$, we obtain a bipartite graph that consists of 554 components of which only one is not a singleton. This particular component has 818 vertices (607 users and 211 performances). Again, we omit the singletons by restricting U to U_D as well as H to H_D and define the *deliberation graph*: $\mathcal{D} = (U_D \cup H_D, E_D, w)$. In this graph an edge represents that a user was at least moderately interested in going to a performance.

Finally, the third graph we want to construct using the given data is the *attendance graph* inherited from the DB_A. We use the edge set $E_A \subseteq U \times H$ defined by $(u,h) \in E_A$ iff u attended performance h. The only non-singleton component has 245 vertices (111 users and 134 performances). By restricting U and H to U_A and H_A we define the bipartite attendance graph as follows: $\mathcal{A} = (U_A \cup H_A, E_A)$.

All those graphs were thus constructed using data provided by the users. They represent various kinds of user interactions with performances. This is not unlike, e. g., users of the twitter platform[1] interacting with tweets, for which some social network character was shown, cf., [12]. Hence, below we target the naturally emerging question, i. e., whether the constructed graphs have the social network property.

4. Social Network Graph

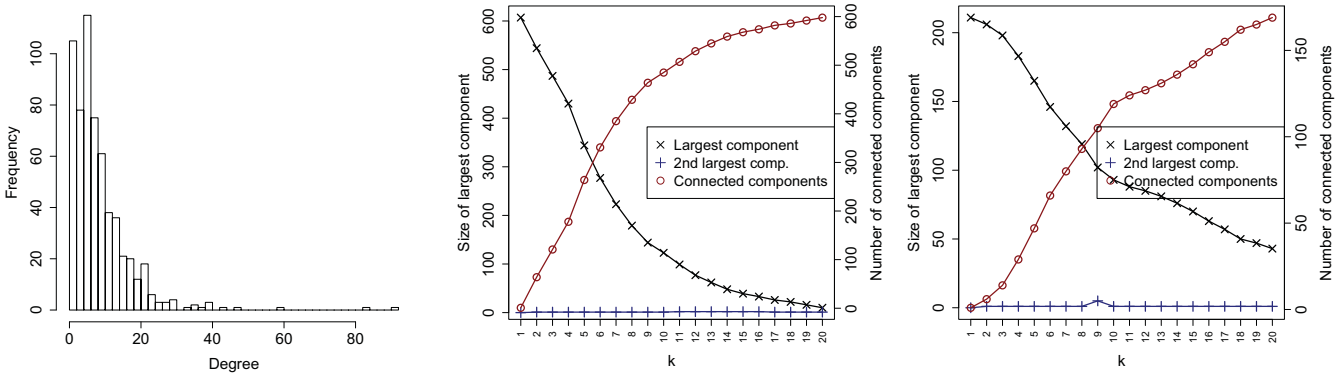
In this section, we analyse if the social event network can be treated as a social network as defined in [13]. For that, we need to show in particular that projections of the just constructed bipartite graphs have characterising values in terms of average path lengths and average local clustering coefficients. If so, this would enable researchers to apply tools and theories developed for such networks. Even more, so far undiscovered differences in social networks could be revealed by comparing this alleged social network to others.

At first we take a closer look at the bipartite graphs themselves. Thereafter, we will project the bipartite graphs onto their user sets, to show that the required defining properties for a social network, as stated in [13], are present.

4.1. Bipartite properties

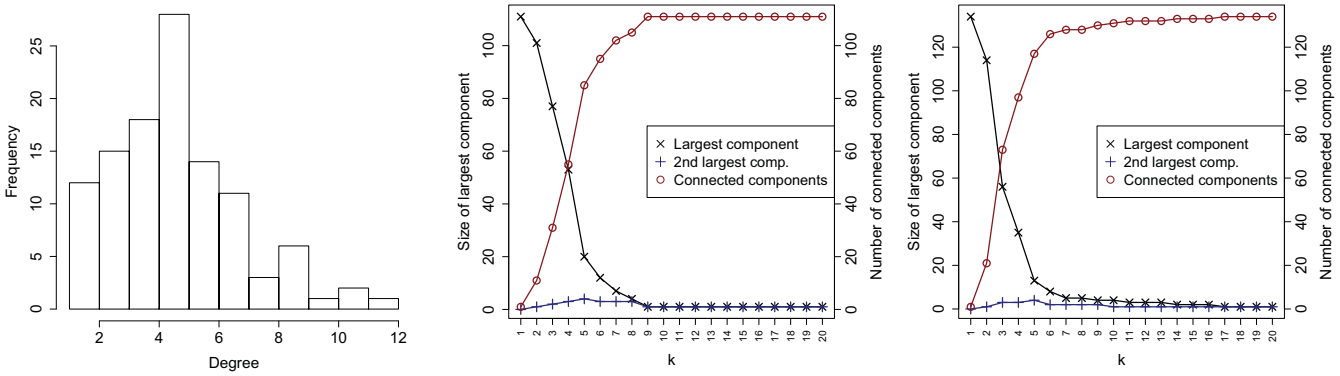
In the following, we study structural properties of the deliberation and the attendance graphs. In particular, we aim at providing an indication about the macroscopic connectivity structure in the bipartite graphs, e. g., regarding characteristic numbers of user–performance connections. For that, we extend methods for visualizing the *k-neighborhood-connectivity* (KNC) [14] for a bipartite graph. Given a bipartite graph $G = (U \cup H, E)$ with a set of vertices U and H and edges E, two vertices in U are *k-neighbours* if there are at least k distinct paths of length two between them (analogously for H). A *k-neighborhood* graph is then induced on U (or H) by this *k-neighborhood*-relation. Then, the KNC-plot shows the degradation of connectivity with increasing k.

1. www.twitter.com



(a) $\mathcal{D}$: user degree distribution (b) $\mathcal{D}$: user KNC-plot (c) $\mathcal{D}$: performance KNC-plot

Figure 1: User degree distribution and KNC-plots for the deliberation graph $\mathcal{D}$: We show user and performance connectivity, plotting the size of the largest and second largest component, respectively, and the number of connected components in the induced KNC-graphs, respectively. In the user KNC-plot, we observe that with an increasing threshold on jointly visited performances, the number of components increases strongly from the start: However, there is only one large component, while the remaining components have a size between 1 and 2. A joint number of five performances still captures about 50% of the total set of users which shows the strong connectivity. We observe similar findings for the performance KNC-plot: Until 7 common users, the largest performance component captures about 50% of the total set of performances; here the remaining components are always at most singletons.



(a) $\mathcal{A}$: user degree distribution (b) $\mathcal{A}$: user KNC-plot (c) $\mathcal{A}$: performance KNC-plot

Figure 2: User degree distribution and KNC-plots for the attendance graph $\mathcal{A}$: We show user and performance connectivity, plotting the size of the largest and second largest component, respectively, and the number of connected components in the induced KNC-graphs, respectively. In the User-KNC-Plot, we observe that with an increasing threshold on jointly visited performances, the number of components increases, even more strongly than in the deliberation graph $\mathcal{D}$. There is also only one large component, while the remaining components have a relatively small size. A joint number of four performances still captures about 50% of the total set of users which shows the strong connectivity. For the performance KNC-plot, until 3 common users, the largest performance component captures about 50% of the total set of performances; the remaining components are relatively small similar to the user KNC-plot.

The original KNC-plot contains the number of connected components as well as the size of the largest component. In addition, we also plot the size of the second largest component in order to obtain a more comprehensive view on the component structure, i. e., for studying whether the split for larger values of k occurs uniformly or not.

Figures 1 and 2 show the degree distributions of the users, and KNC-plots for users and performances, respectively. In the deliberation graph, we observe a skewed distribution, which however, conforms to the (behavioural) trends considering the degree distribution of the attendance graph.

On average about five performances are visited – this is also reflected by the number of performances assigned with preferences in the deliberation graph. Further, we observe a strong connectivity structure both for the deliberation and the attendance graph. With increasing k the largest component breaks up only gradually. This already indicates interesting socio-spatial characteristics towards those of a social network, since we can assume a strong (local) clustering structure. We will analyse that in more detail in Section 4.2. Furthermore, we observe a "meaningful" number of performances for a user up to five performances for that social structure.

TABLE 1: Quantitative properties of the projected graphs.

Graph	Vertices	Edges	Edge-density
$\mathcal{P}_U$	612	78415	0.42
$\mathcal{D}_U$	607	69198	0.38
$\mathcal{A}_U$	111	1145	0.19
$\mathcal{D}_U^*$	79	1341	0.44
$\mathcal{A}_U^*$	78	586	.20

4.2. User-Projections

We project all mentioned bipartite graphs on their set of users U to obtain simple undirected graphs, cf., Table 1 for an overview. This approach is common to analyse bipartite networks for social network properties, cf., actor collaboration in [13]. For $\mathcal{P}$ we construct $\mathcal{P}_U \coloneqq (U_P, E_P^U)$ with $E_P^U \coloneqq \{(u,v) \in U_P \mid \exists h \in H\colon (u,h) \in E_P \wedge (v,h) \in E_P\}$, for $\mathcal{D}_U$ and $\mathcal{A}_U$ analogously. The analysis would be more straightforward if the sets U_A and U_D were identical. However, this is not the case. In fact the intersection of those user sets has 79 users. The restrictions of the projections of the deliberation graph as well as of the attendance graph to those 79 users, are indicated by $\mathcal{D}_U^*$ and $\mathcal{A}_U^*$ respectively[2].

Watts stated in [15] that social network like graphs have specific characteristics in terms of local clustering and global separation, cf., [16] for a comparison. This also holds for graph projections from a social bipartite graph [17].

Graph properties. Social networks in general show the small-world effect, i. e., small average shortest path lengths [15], and high average local clustering coefficients [13]. Many observations of network properties can be explained just by the network's degree distribution [18]. It is therefore important to contrast the observed small-world properties to the according results obtained from a random graph (i. e., a *null model*) sharing the same degree distribution. To obtain such null models for our graphs we use the algorithm from [19], which shuffles the edges of a given graph G preserving the degree of every vertex, i. e., the number of edges intersecting with the vertex. This process is typically repeated a multiple of the graph edge set's cardinality [20]. In our experiments we shuffled as often as 100 times the cardinality of the edge set.

Average Shortest Path Length. A *path* $v_0 \to_G v_n$ of *length* n in a graph $G = (V, E)$ is a vertex sequence $(v_0, \ldots, v_n) \in V^{n+1}$, $n \geq 1$ and $\{v_i, v_{i+1}\} \in E$ for all $i = 0, \ldots, n-1$. A *shortest path* between nodes u and v is a path $u \to_G v$ of minimal length.

The average shortest path length for a social network, i. e., the mean of shortest path lengths for any two vertices in a graph, is significantly low for social network graphs. The follower graph of the social network twitter, for example, has an average path length of 4.17, see [12]. A data set which is even more comparable to our investigated data, but rather small, is the southern woman data set [21]; it recorded the participation of 14 persons to 18 events. Here, the average shortest path length is 1.09. In contrast, the Internet router network [22] has an average shortest path length of 9.51.

2. $\mathcal{A}_U^*$ has two connected components, one with 78 vertices, the other one being a singleton only. We ignore the latter in the subsequent analyses.

TABLE 2: Average shortest path lengths (ASP) of our projected graphs compared to comparable Watts-Strogatz-Graphs with $p = 0$ ($WSG0$) and $p = 0.1$ ($WSG1$), random graph R, and null model (NM).

Graph	ASP	$WSG0$	$WSG1$	R	NM
$\mathcal{P}_U$	1.58	1.74	1.58	1.58	1.58
$\mathcal{D}_U$	1.63	1.87	1.62	1.62	1.62
$\mathcal{A}_U$	2.02	3.27	2.12	1.82	1.92
$\mathcal{D}_U^*$	1.58	1.69	1.56	1.56	1.58
$\mathcal{A}_U^*$	2.05	2.92	2.07	1.84	1.99

For the investigated projections we obtained low average shortest path lengths, between 1.58 for $\mathcal{P}_U$ and 2.05 for $\mathcal{A}_U^*$, see Table 2. For every graph we compute the following list of graphs and compare their properties with original one: the Watts-Strogatz [13] model (WSG) using $p = 0.0$ and $p = 0.1$, a random graph with the same amount of vertices and edges, and the null model. For the projections of $\mathcal{P}$ and $\mathcal{D}$ we observed almost identical values for WSG in average shortest lengths for $p = 0.1$. The other projections seem to be closely reproduced in terms of average shortest path. The null model behaves alike but the discrepancy for $\mathcal{A}_U$ and $\mathcal{A}_U^*$ is distinct. So the graphs that emerged from the initial data sets, the graphs constructed by the Watts-Strogatz model, and the null model graph have no distinct demeanour in terms of average shortest path length for graphs of this size. Given these results we may claim that this definitional requirement for a small-world network, which is necessary to qualify as a social graph, is satisfied. However, we need to check the other requirement for small-world networks to substantiate our initial assumption.

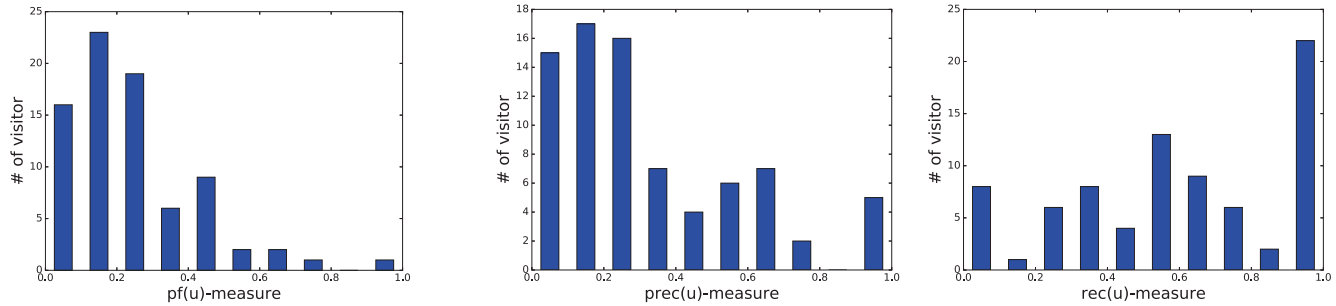
Average local clustering coefficient. Small-world network graphs tend to have a high average local clustering coefficient ($alcc$), see [13]. That is $\frac{1}{n} \cdot \sum_{i=1}^{n} C_i$ where C_i is the local clustering coefficient for the vertex v_i. This coefficient can be computed as follows. Let $N_i \coloneqq \{v \in V \mid \{v_i, v\} \in E\}$, the neighbourhood of v_i. We then compute:

$$C_i = \frac{2|\{e_{ij} \mid v_i, v_j \in N_i, e_{ij} \in E\}|}{|N_i|(|N_i| - 1)}$$

For example, the aforementioned Internet router network has an $alcc$ of 0.03, see [13], the graph of twitter followers has an $alcc$ of 0.3 (see [12]) and the social network formed by actors has an $alcc$ of 0.79, see [13].

Table 3 presents the clustering coefficients for all projections, the Watts-Strogatz model, a pure random graph, and the null model. For all graphs emerging from DB_P and DB_A we find a high $alcc$. Like [5], we note that $alcc$ is smaller for offline than online networks. We further observe the necessary property that for all projections the $alcc$ obtained from the degree-preserving randomised graphs decreases. Especially for $\mathcal{A}_U$ and $\mathcal{A}_U^*$ this decrease is very high.

Summary. Our results concerning the average shortest path length combined with the results for the clustering coefficient underpin the claim that the LNM event can be treated as a social network. Hence, the investigation of the LNM dataset in the realm of social networks is meaningful.



(a) Histogram for $pf(u)$ for all 79 users in U_B.

(b) Histogram for the pre and rec measure for all 79 users in U_B.

Figure 3: Histograms for the user-based plan fulfilment measures, for all 79 users (U_B) where deliberation and attendance is known.

TABLE 3: Clustering coefficient (CC) of our projected graphs compared to comparable Watts-Strogatz-Graphs with $p = 0$ ($WSG0$) and $p = 0.1$ ($WSG1$), random graph R, and null model (NM) .

Graph	CC	$WSG0$	$WSG1$	R	NM
$\mathcal{P}_U$	0.75	0.75	0.60	0.42	0.70
$\mathcal{D}_U$	0.71	0.75	0.60	0.38	0.63
$\mathcal{A}_U$	0.52	0.71	0.54	0.19	0.30
$\mathcal{D}^*_U$	0.74	0.73	0.60	0.43	0.69
$\mathcal{A}^*_U$	0.49	0.70	0.52	0.19	0.31

5. Behaviour Analysis

This section investigates the deliberation vs. the attendance graph, i. e., to what extent the deliberation determines the attendance, and proposes suitable measures. Further, we introduce newly developed methods to discover particular interesting groups of users and groups of performances.

5.1. Plan Fulfilment: Intention vs. Reality

For comparing $\mathcal{A}$ and $\mathcal{D}$, we restrict the set of users U to users where both, attendance and deliberation, is known. This is true for 79 users. We denote this set by U_B. Using both graphs we use the Jaccard-Distance to compare the deliberation to the actual attendance of a user $u \in U_B$. For that let

$$pf(u) \coloneqq \frac{|\,N_A(u) \cap N_D(u)|}{|\,N_A(u) \cup N_D(u)|},$$

where $N_D = \{h \in V \mid (u,h) \in E_D\}$, i. e., the set of performances u was planning to go to, and $N_A = \{h \in V \mid (u,h) \in E_A\}$, i. e., the set of performances u actually went to. This measure does not take into account that there are different levels of deliberation to a performance. Therefore, we call it *simple plan follow measure*. A histogram for this measure using ten bins is shown in Figure 3a. We find the number of visitors that actually completely fulfilled their plan is one. The number of visitors with a fulfilment of up to 0.5 is 73. But only 8 users did have a $pf(u)$ of zero, i. e., they did not attend any of the planned performances. However, as we looked them up we found that they in fact attended between one and nine performances. The average fulfilment for all users is 0.23; this (low) number can be partially attributed to the tour planning app for generating a plan from a (larger) set of preferences. Therefore, since the $pf(u)$-measure does not explain if the reason for a low value is due to participating in fewer performances or to participating in not planned performances. To answer this striking question we adapt two measures known from information retrieval, i. e., precision (pre) and recall (rec):

$$\mathrm{pre}(u) \coloneqq \frac{|\,N_D(u) \cap N_A(u)|}{|\,N_D(u)|}, \; \mathrm{rec}(u) \coloneqq \frac{|\,N_A(u) \cap N_D(u)|}{|\,N_A(u)|}$$

The rec-measures yields 1.0 for a user u if she attended only intended performances; the pre-measure yields 1.0 for a user u if she attended all performances she had planned to go to. From the plots shown in Figure 3b we may conclude the following statements for LNM. The majority of users attended to performances they planned beforehand to some extent. In particular, for a lower recall bound in rec of 0.6 there are 39 users. Further, the vast majority did not go where they planned to. Only 20 users have a precision of at least 0.5, i. e., went at least to half of the planned performances.

5.2. Analysis of Group Individuality

A clique, i.e., a vertex subset of an undirected (bipartite) graph such that its induced subgraph is edge-complete, is an object of research for group detection for about 70 years [23]. During the early 1980s, Wille and Ganter developed a theory of data analysis suitable for bipartite graph-like data that is able to order maximal cliques, called *formal concept analysis* (FCA), see [1]. Here, maximal cliques are regarded as formal concepts, which form a partially ordered set: For each pair of elements this set has a unique least upper bound and a unique greatest lower bound, characterizing the mathematical structure of a lattice. One of the first lattice-based investigations to represent a social network and detect groups in it was done by Freeman [23], [24]. Below, we characterise the user group as a whole and depict particular interesting subsets of users in terms of performance participation behaviour.

We provide a very brief introduction to FCA. The analysis is based on a *formal context* that is a triple $\mathbb{K} = (G, M, I)$ which consists of an object set G, an attribute set M and an incidence relation $I \subseteq G \times M$. We say that an object $g \in G$ has an attribute $m \in M$ iff $(g, m) \in I$. The formal

TABLE 4: Formal context example taken from DB_P.

$\mathbb{K}$	Carl Orff	Wild Society	Cafe Camera	Caribbean
userA	X	X		
userB	X		X	X
userC	X			X
userD				X

context is often represented utilising a cross table, see Table 4 for an example. A *formal concept* then is a pair (X, Y) with $X \subseteq G$, $Y \subseteq M$, such that $X' = Y$ and $Y' = X$, where $X' := \{m \in M \mid (x, m) \in I \text{ for all } x \in X\}$ and $Y' := \{g \in G \mid (g, y) \in I \text{ for all } y \in Y\}$. For example, for the context shown in Table 4, a formal concept would be $(\{\text{userB}, \text{userC}\}, \{\text{Carl Orff}, \text{Caribean}\})$. In words, the maximal set of attributes userB and userC have in common is the attribute set containing Carl Orff and Caribean. On the other hand, the maximal set of objects the attributes Carl Orff and Caribbean have in common is the object set containing userB and userC. Another formal concept in $\mathbb{K}$ from Table 4 would be $(\{\text{userA}\}, \{\text{Carl Orff}, \text{Wild Society}\})$. We denote with $\mathfrak{B}(\mathbb{K})$ the set of all formal concepts emerging from the formal context $\mathbb{K}$. We may order formal concepts using $(X_1, Y_1) \leq (X_2, Y_2) :\Leftrightarrow X_1 \subseteq X_2$ for all $X_1, X_2 \subseteq G$ and $Y_1, Y_2 \subseteq M$. The ordered set $\mathfrak{B}(\mathbb{K})$ is a complete lattice, called the concept lattice of $\mathbb{K}$.

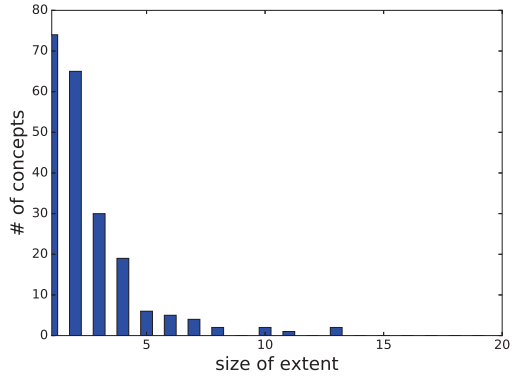


Figure 4: Number of concepts for particular sizes of extents for $\mathbb{K}$.

We present some insights into the behaviour of the users that FCA can provide very easily. For that, we apply FCA to the attendance graph restricted to the user set U_B. The formal context $\mathbb{K}_{\text{LNM}} = (U_B, H_A, E_A \cap (U_B \times H_A))$ results in 79 objects (the users) and 121 attributes (performances). The associated concept lattice has 212 formal concepts. A visualisation of a concept lattice this size would be unhelpful. We therefore want to analyse properties of the obtained concepts. In Figure 4 we plotted a histogram for the number of formal concepts according to the size of its extent, i. e., the number of users. The biggest column is for extent size one, in particular there are 74 concepts. By this we learn that there are 74 unique combinations of performances which were realised by those 74 individuals during LNM. Obviously, the number of users is an upper bound for formal concepts with extent size one. This fact can be used intuitively to informally define a coefficient for the user individuality of a distributed social event as follows. Let $\mathbb{K}$ be the formal context of a social event network graph defined in the way above, then the *user individuality measure* is:

$$\text{uic}(\mathbb{K}) = \frac{|\{(X, Y) \in \mathfrak{B}(\mathbb{K}) \mid |X| = 1\}|}{|U_B|}$$

In events where there is no concept of extent size one, this coefficient would return zero. Then again, in events where there are as many concepts with extents of size one as there are users, the uic would yield 1.0. In the case of LNM this coefficient is high: $\text{uic}(\mathbb{K}_{\text{LNM}}) = 0.93$. This indicates a high level of individuality among the users. For example, if we apply the same analysis to the event based southern woman social network graph we get: $\text{uic}(\mathbb{K}_{\text{southern}}) = 0.22$.

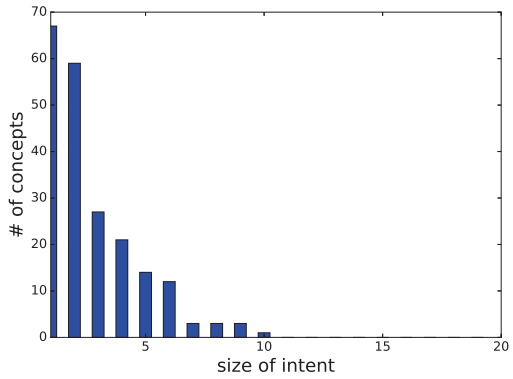


Figure 5: Number of concepts for particular sizes of intents for $\mathbb{K}$.

We may apply a similar analysis for the intents as well, see Figure 5: We observe 67 formal concepts with an intent size of one. The maximal number can here be determined by the size of the attribute set, i. e., 212. Based on that we want to propose as performance individuality coefficient:

$$\text{pic}(\mathbb{K}) = \frac{|\{(X, Y) \in \mathfrak{B}(\mathbb{K}) \mid |Y| = 1\}|}{|H_A|}$$

One might be easily misled by the thought that uic somehow determines pic or vice versa. This thought can be refuted by the following construction. An event may have 231 performances and 100 users. Let 22 of the 100 users share precisely one performance, which is possible since there are $\binom{22}{2}$ possible combinations. From the remaining 78 users take again a set of 22 users from which each user copies the behaviour from exactly one user of the previous 22. The rest of 56 users may copy the behaviour of some user from the first 22. The constructed event network has a high pic since all performances are attended by a unique group of users. However, the uic would be zero since no user with a unique set of attended performances is present.

The larger $\text{pic}(\mathbb{K}_{\text{LNM}})$, the more performances exist with a unique set of participants, and the more individual the whole event is. For LNM $\text{pic}(\mathbb{K}_{\text{LNM}}) = 0.31$, compared to the southern woman social network where this coefficient is 0.28. So when looking on the event through the performances, the event was not as individual as uic might indicate.

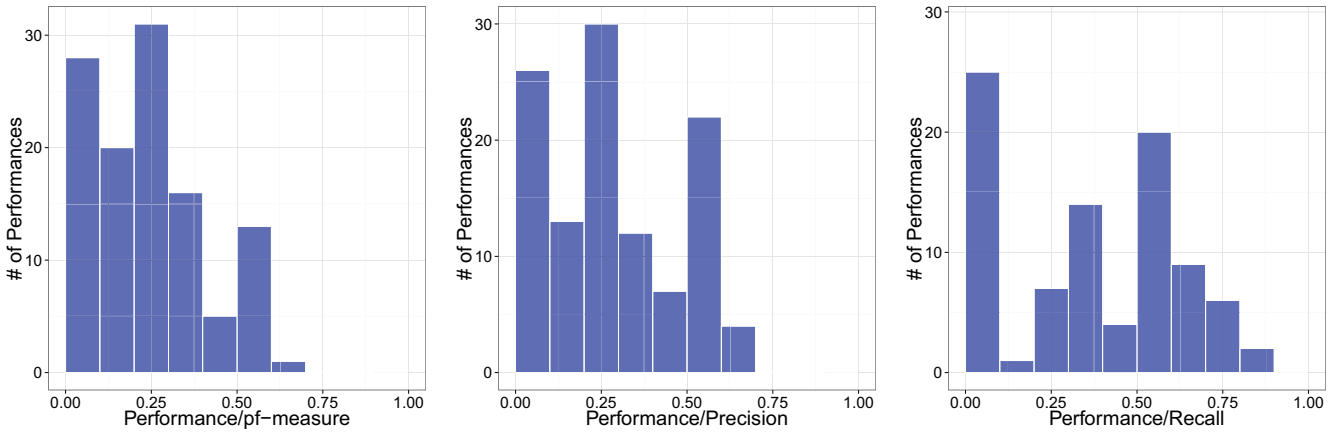


Figure 6: Histograms for plan fulfilment pf, and the precision pre and recall rec measure adaptations with respect to performances.

These numbers reflect the intuition that different 'interesting' performances might be attended by the same set of users, whereas all of those users might have very deviant attendance for the evening. We suppose that pic is low for social networks in general since the properties of LNM can be mapped to properties of other social networks.

5.3. Characterising Performance Fulfilment

In the following, we first focus on the plan fulfilment from the performances' perspective. After that, we characterise distinct subgroups of performances utilizing these measures.

5.3.1. Performance Fulfilment. We restrict our analysis to the set of performances that is observed both in the deliberation and the attendance graph, resulting in 117 performances. Analogously to the *plan fulfilment* measures of a user (see above), we can define according measures with respect to specific performances, e. g., the plan fulfilment $pf(h)$ for a performance h as:

$$pf(h) := \frac{|\,N_A(h) \cap N_D(h)|}{|\,N_A(h) \cup N_D(h)|},$$

where $N_D = \{u \in V \mid (u,h) \in E_D\}$, i. e., the set of users that planned to attend h, and $N_A = \{u \in V \mid (u,h) \in E_A\}$, i. e., the set of users that actually attended to h.

Accordingly, we define a precision measure ($\mathrm{pre}(h)$) and recall measure ($\mathrm{rec}(h)$):

$$\mathrm{pre}(h) := \frac{|\,N_D(h) \cap N_A(h)|}{|\,N_D(h)|}, \; \mathrm{rec}(h) := \frac{|\,N_A(h) \cap N_D(h)|}{|\,N_A(h)|}$$

Intuitively, the maximal $\mathrm{pre}(h)$ value of 1.0 implies that all users that attended h also had the intention to visit h, while the maximal $\mathrm{rec}(h)$ value of 1.0 implies that all users that intended to attend h really attended h. Conversely, the minimal values of 0.0 indicate that there is a complete mismatch between intention and reality concerning the set of users of a specific performance h.

Averaging over all performances, we obtain the mean values $\mu_0(pf) = 0.24$, $\mu_0(\mathrm{pre}) = 0.28$, $\mu_0(\mathrm{rec}) = 0.51$. Figure 6 shows the respective distributions of these measures. Essentially, these result support our findings discussed in Section 5.1. However, they allow a fine-grained analysis from the perspective of performances. We observe that *complete* plan fulfilment is rare. Instead, users tended to enter a variety of preferences, while they also tended to attend a significant portion of their entered preferences. Nevertheless, we also observe a large number of performances that were visited in ad-hoc fashion ($\mathrm{pre}(h) = \mathrm{rec}(h) = 0$, for the specific performance h).

5.3.2. Characterising Performance Subgroups. In order to obtain a more detailed view on performance fulfilment, we applied subgroup discovery in order to identify groups of performances that exhibit a large (and deviating) value of the pf, pre, or rec measure. Subgroup discovery, e. g., [2], aims at identifying subgroups of individuals that are *interesting* with respect to a certain target concept. For a numerical target concept like pf, for example, we are interested in large subgroups with a high share of individuals for which the target concept deviates from the total population, e. g., estimated by comparing the respective mean values. Then, we aim at discovering *subgroup descriptions* that are made up by conjunctions of features (selection expressions), e. g., by *handmade AND music*, or *hit AND rock* indicating interesting groups of performances. The subgroups are then the groups of performances covered by the respective subgroup description.

For characterising these groups, we extracted descriptive information from the textual information of the event: This included *genre* categories as well as descriptions (free text) of the individual performances. We applied typical data preprocessing steps such as stemming and stop word removal, e. g., [25]. We also filtered words below a minimal frequency threshold $\tau = 5$ reducing a total number of 2767 to 180 descriptive words (features). For estimating the interestingness of a subgroup S, we applied the simple binomial quality function [2]: $q(S) = \sqrt{n} \cdot (\mu_S(t) - \mu_0(t))$, where $\mu_S(t)$ and $\mu_0(t)$ denote the mean values of the target concept t in the subgroup and in the total dataset, respectively, and n indicates the size of the subgroup.

TABLE 5: Top subgroups w.r.t. measures of performance fulfilment.

#	$\mu(pf)$	$\mu(\text{pre})$	$\mu(\text{rec})$	Size	Description
1	**0.8**	**0.78**	**0.67**	5	*handmade AND music*
2	**0.73**	**0.64**	**0.79**	5	*hit AND rock*
3	0.31	0.32	**0.93**	7	*traditional*

Table 5 indicates the top subgroups with respect to different performance fulfilment measures. Groups #1 and #2 are the top-2 subgroups for the sum of all three measures indicating a good overall plan fulfilment, i. e., a good match between intention and reality. Subgroup #3 is the top subgroup for recall: Performances described by *traditional* score well only for that measure. This indicates, that here mostly participants (with an interest in traditional music) attended that had also the intention to attend.

6. Conclusion

In this paper, we focused on the analysis of the social structure of a distributed event (LNM) considering preferences and visiting behaviour of the users. We assumed that these can be regarded as expressing socio-spatial characteristics, and analysed the corresponding bimodal network of user–performance relations. Our novel results concerning this type of data show, that this social event network conforms to the characteristics of a social network. In addition, we demonstrated that we can identify distinct behavioural patterns concerning groups of users and performances, respectively. We focused on the attendance behaviour of events and the match between deliberation (intention) and attendance for a certain performance. Performing our analysis, we propose and demonstrate a set of methods for the analysis of such network models: We applied the extended KNC-plot method, and novel group description methods based on Formal Concept Analysis for identifying and describing characteristic user groups. Furthermore, we analysed plan fulfilment concerning performances and their associated user groups using subgroup discovery. Altogether, our results indicate strong connectivity of the users and performances (on common sets of performances and users, respectively), while we also observe strong individuality of the users at the same time. This indicates a certain *common core* behaviour, which is further adapted according to individual interests.

For future work, we aim to study more networks on distributed events, also integrating information from online social networks. Furthermore, we plan to integrate our findings into recommendation algorithms, and extend the analysis, since the results on plan fulfilment provide indications on necessary assistive functionality concerning sequences of performances and the flexibility of plan adaptation. Also, by utilizing the results on user and event (sub-)groups the development of according personalization approaches (and their analysis) is another interesting direction for future work.

References

[1] B. Ganter and R. Wille, *Formal Concept Analysis: Mathematical Foundations*. Berlin/Heidelberg: Springer, 1999.

[2] M. Atzmueller, "Subgroup Discovery – Advanced Review," *WIREs: DMKD*, vol. 5, no. 1, pp. 35–49, 2015.

[3] S. Scellato, A. Noulas, R. Lambiotte, and C. Mascolo, "Socio-Spatial Properties of Online Location-Based Social Networks," in *Proc. ICWSM*. Palo Alto, CA, USA: AAAI Press, 2011, pp. 329–336.

[4] M. Atzmueller, S. Doerfel, A. Hotho, F. Mitzlaff, and G. Stumme, "Face-to-Face Contacts at a Conference: Dynamics of Communities and Roles," in *Modeling and Mining Ubiquitous Social Media*, ser. LNAI, 2012, vol. 7472.

[5] X. Liu, Q. He, Y. Tian, W.-C. Lee, J. McPherson, and J. Han, "Event-based Social Networks: Linking the Online and Offline Social Worlds," in *Proc. SIGKDD*. New York, NY, USA: ACM, 2012, pp. 1032–1040.

[6] M. Atzmueller, "Data Mining on Social Interaction Networks," *Journal of Data Mining and Digital Humanities*, vol. 1, June 2014.

[7] B.-E. Macek, C. Scholz, M. Atzmueller, and G. Stumme, "Anatomy of a Conference," in *Proc. Hypertext*. ACM, 2012, pp. 245–254.

[8] F. Mitzlaff, M. Atzmueller, A. Hotho, and G. Stumme, "The Social Distributional Hypothesis," *SNAM*, vol. 4, no. 216, 2014.

[9] R. Schaller, M. Harvey, and D. Elsweiler, *Advances in Information Retrieval. Proc. ECIR*, 2014, ch. Detecting Event Visits in Urban Areas via Smartphone GPS Data.

[10] R. Schaller, "Electronic Tourist Guides: User-friendly Editing of Automatically Planned Routes," in *Proc. STAIRS 2014*, ser. Frontiers in Artificial Intelligence and Applications, pp. 260–269.

[11] R. Schaller, M. Harvey, and D. Elsweiler, "RecSys for Distributed Events: Investigating the Influence of Recommendations on Visitor Plans," in *Proc. SIGIR*. ACM, 7 2013.

[12] S. A. Myers, A. Sharma, P. Gupta, and J. Lin, "Information Network or Social Network?: The Structure of the Twitter Follow Graph," in *Proc. WWW (Companion)*. NY, NY, USA: ACM, 2014, pp. 493–498.

[13] D. J. Watts and S. H. Strogatz, "Collective Dynamics of /'Small-World/' Networks," *Nature*, vol. 393, no. 6684, pp. 440–442, 1998.

[14] R. Kumar, A. Tomkins, and E. Vee, "Connectivity Structure of Bipartite Graphs via the KNC-Plot," in *Proc. WSDM*, New York, NY, USA, 2008, pp. 129–138.

[15] D. J. Watts, "Networks, Dynamics, and the Small-World Phenomenon," *American Journal of Sociology*, vol. 105, pp. 493–527, 1999.

[16] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D.-U. Hwang, "Complex Networks: Structure and Dynamics," *Physics Reports*, vol. 424, no. 45, pp. 175 – 308, 2006.

[17] J.-L. Guillaume and M. Latapy, "Bipartite Graphs as Models of Complex Networks," *Physica A: Statistical Mechanics and its Applications*, vol. 371, no. 2, pp. 795 – 813, 2006.

[18] E. Kolaczyk, "Statistical analysis of network data: Methods and models," *Springer Series In Statistics*, p. 386, 2009.

[19] C. G. M. Mihail and E. Zegura, "The Markov Chain Simulation Method for Generating Connected Power Law Random Graphs," in *Proc. 5th Workshop on Algorithm Engineering and Experiments*, vol. 111. SIAM, 2003, p. 16.

[20] S. Maslov and K. Sneppen, "Specificity and stability in topology of protein networks," *Science*, vol. 296, no. 5569, p. 910, 2002.

[21] L. C. Freeman, "Finding Social Groups: A Meta-Analysis of the Southern Women Data," in *Dynamic Social Network Modeling and Analysis: Workshop Summary and Papers (pp 39-97), Washington D.C.: National Research Council, The National Academies*, 2002.

[22] A. Vázquez, R. Pastor-Satorras, and A. Vespignani, "Internet Topology at the Router and Autonomous System Level," *CoRR*, vol. cond-mat/0206084, 2002.

[23] L. C. Freeman, "Cliques, galois lattices, and the structure of human social groups," *Social Networks*, vol. 18, no. 3, pp. 173 – 187, 1996.

[24] L. Freeman and D. White, "Using galois lattices to represent network data," *Sociological Methodology*, vol. 23, pp. 127–146, 1993.

[25] A. Schmidt, M. Atzmueller, and M. Hollender, "Data Preparation for Big Data Analytics: Methods & Experiences," in *Enterprise Big Data Engineering, Analytics, and Management*. IGI Global, 2016.