Compositional Subgroup Discovery on Attributed Social Interaction Networks

Martin Atzmueller

Tilburg University, Department of Cognitive Science and Artificial Intelligence,
Warandelaan 2, 5037 AB Tilburg, The Netherlands
m.atzmuller@uvt.nl

Abstract. While standard methods for detecting subgroups on plain social networks focus on the network structure, attributed social networks allow compositional analysis, i. e., by exploiting attributive information. Thus, this paper applies a compositional perspective for identifying compositional subgroup patterns. In contrast to typical approaches for community detection and graph clustering, it focuses on the dyadic structure of social interaction networks. For that, we adapt principles of subgroup discovery – a general data mining technique for the identification of local patterns – to the dyadic network setting. We focus on social interaction networks, where we specifically consider properties of those social interactions, i. e., duration and frequency. In particular, we present novel quality functions for estimating the interestingness of a subgroup and discuss their properties. Furthermore, we demonstrate the efficacy of the approach using two real-world datasets on face-to-face interactions.

1 Introduction

The identification of interesting subgroups (often also called communities) is a prominent research direction in data mining and (social) network analysis, e. g., [2, 3, 17, 21, 49]. Typically, a structural perspective is taken, such that specific subgraphs — in a graph representation of the network — induced by a set of edges and/or nodes are investigated. Attributed networks, where nodes and/or edges are labeled with additional information, allow further dimensions for detecting patterns that describe a specific subset of nodes of the graph representation of a (social) network. However, there are different foci relating to the specific problem and data at hand. The method of subgroup discovery, for example, a powerful and versatile method for exploratory data mining, focuses on detecting subgroups described by specific patterns that are interesting with respect to some target concept and quality function. In contrast, community detection, as a (social) network analysis method, aims at detecting subgroups of individuals, i. e., nodes of a network, that are densely (and often cohesively) connected by a set of links. Thus, the former stresses the compositional notion of a pattern describing a subgroup, i. e., based on attributes/properties of nodes and/or edges, while the latter focuses on structural properties of a pattern, such that specific subgraphs are investigated that induce a specific pattern.

Preprint of: Martin Atzmueller. Compositional Subgroup Discovery on Attributed Social Interaction Networks. In: Proc. International Conference on Discovery Science. Springer, 2018

Problem. We formalize the problem of detecting compositional patterns of actor-dyads, i. e., edges connecting two nodes (corresponding to the actors) in a graph representation of an attributed network. We aim to detect the subgroup patterns that are most interesting according to a given interestingness measure. For estimating the interestingness, we utilize a quality function which considers the dyadic structure of the set of dyads induced by the compositional pattern. In particular, we focus on social interaction networks, where we specifically consider properties of social interactions, e. g., duration and frequency. Then, the quality measure should consider those patterns as especially interesting which deviate from the expected "overall" behavior given by a null-model, i. e., modeling dyadic interactions due to pure chance. Then, those models should also incorporate the properties of social interaction networks mentioned above.

Objectives. We tackle the problem of detecting compositional patterns capturing subgroups of nodes that show an interesting behavior according to their dyadic structure as estimated by a quality measure. We present novel approaches utilizing subgroup discovery and exceptional model mining techniques [3, 7, 18]. Further, we discuss estimation methods for ranking interesting patterns, and we propose two novel quality functions, that are statistically well-founded. This provides for a comprehensive and easily interpretable approach for this problem.

Approach & Methods. For our compositional subgroup discovery approach, we adapt principles of subgroup discovery – a general data analysis technique for exploratory data mining – to the dyadic network setting. In particular, we present two novel quality functions for estimating the interestingness of a subgroup and its specific dyadic interactions and discuss their properties. Furthermore, we demonstrate the efficacy of the approach using two real-world datasets.

Contributions. Our contribution is summarized as follows:

1. We formalize the problem of compositional subgroup discovery and present an approach for detecting compositional subgroup patterns capturing interesting subgroups of dyads, as estimated by a quality function.
2. Based on subgroup discovery and exceptional model mining techniques, we propose a flexible modeling and analysis approach, and present two novel interestingness measures for compositional analysis, i. e., quality functions for subgroup discovery. These enable estimating the quality of subgroup patterns in order to generate a ranking. The proposed quality functions are statistically well-founded, and provide a statistical significance value directly, also easing interpretation by domain specialists.
3. We demonstrate the efficacy of our proposed approach and the presented quality measures using two real-world datasets capturing social face-to-face interaction networks.

Structure. The rest of the paper is structured as follows: Section 2 discusses related work. After that, Section 4 outlines the proposed approach. Next, Section 5 presents results of an exploratory analysis utilizing two real-world social interaction network datasets of face-to-face interactions. Finally, Section 6 concludes with a discussion and interesting directions for future work.

2 Related Work

Below, we summarize related work on subgroup discovery, social interaction networks, and community detection, and put our proposed approach into context.

2.1 Subgroup Discovery and Exceptional Model Mining

Subgroup discovery is an exploratory data mining method for detecting interesting subgroups, e. g., [3, 29, 50]. It aims at identifying descriptions of subsets of a dataset that show an interesting behavior with respect to certain interestingness criteria, formalized by a quality function, e. g., [50]. Here, the concept of exceptional model mining has recently been introduced [18, 34]. It can be considered as a variant of subgroup discovery enabling more complex target properties. Applications include mining characteristic patterns [8], mining subgroups of subgraphs [45], or descriptive community mining, e. g., [7]. In contrast to the approaches mentioned above, we adapt subgroup discovery for dyadic analysis on social interaction networks, and propose novel interestingness measures as quality functions on networks for that purpose.

2.2 Mining Social Interaction Networks

A general view on mining social interaction networks is given in [2]. captured during certain events, e. g., during conferences. Here, patterns on face-to-face contact networks as well as evidence networks [40]) and their underlying mechanisms, e. g., concerning homophily [11, 39, 41] are analyzed, however only concerning specific hypotheses or single attributes [46]. Furthermore, [6,38] describe the dynamics of communities and roles at conferences, while [28] focuses on their evolution. This is also the focus of, e. g., [4,37] where exeptional communities/subgroups with respect to sequential transitions are detected. In contrast, this paper targets the detection of interesting patterns describing such dyadic-oriented subgroups in attributed networks, modeling social interactions.

Attributed (or labeled) graphs as richer graph representations enable approaches that specifically exploit the descriptive information of the labels assigned to nodes and/or edges of the graph, in order to detect densely connected groups or clusters, e. g., [16]. In [7], for example, the COMODO algorithm is presented. It applies subgroup discovery techniques for description-oriented community detection. Using additional descriptive features of the nodes contained in the network, the task is to identify communities as sets of densely connected nodes together with a *description*, i. e., a logical formula on the values of the nodes' descriptive features. Here, in contrast, we do not focus on the graph structure, like approaches for community detection, e. g., [7,24,44] or exceptional model mining approaches, e. g., [10, 12, 15, 26] on attributed graphs. Instead, we apply a dyadic perspective on interactions focusing on such parameters such as interaction *frequency* and *duration*. We propose two novel quality functions in such dyadic interaction contexts, i. e., for reliably identifying interesting subsets of dyads using subgroup discovery. To the best of the author's knowledge, no subgroup discovery approach tackling this problem has been proposed so far.

3 Background: Subgroup discovery

Subgroup discovery [3,50] is a powerful method, e. g., for (data) exploration and descriptive induction, i. e., to obtain an overview of the relations between a so-called target concept and a set of explaining features. These features are represented by attribute/value assignments, i. e., they correspond to binary features such as items known from association rule mining [1]. In its simplest case, the target concept is often represented by a binary variable. However, more complex target concepts can also be modeled, leading to exceptional model mining which targets specifically complex target models. In this work, for subgroup discovery we adopt the general scope proposed in [3, 29–31, 36, 43, 50, 51], such that subgroup discovery also contains exceptional model mining as a special case, enabling more complex target concepts than just, e. g., a single dependent variable. Then, subgroups are ranked using a quality function, e. g., [3, 22, 29, 35, 50].

In the context of attributed networks, we formalize the necessary notions in the following. Formally, an *edge – attribute database* $DB = (E, A, F)$ is given by a set of edges E and a set of attributes A. For each attribute $a \in A$, a range $dom(a)$ of values is defined. An attribute/value assignment $a = v$, where $a \in A, v \in dom(a)$, is called a *feature*. We define the feature space V to be the (universal) set of all features. For each edge $e \in E$ there is a mapping $F : E \rightarrow 2^V$ describing the set of features that are assigned to an edge. Intuitively, such features can be given by attribute–value paris, (binary) labels such as items in the context of association rule mining, etc.

Basic elements used in subgroup discovery are patterns and subgroups. Intuitively, a *pattern* describes a *subgroup*, i. e., the subgroup consists of the edges (and the respective nodes) that are covered by the respective pattern, i. e., those having the respective set of features. It is easy to see, that a pattern describes a fixed set of edges (inducing a subgroup of nodes), while a subgroup can also be described by different patterns, if there are different options for covering the subgroup' edges. A (subgroup) *pattern* P is defined as a conjunction

$$P = s_1 \wedge s_2 \wedge \ldots \wedge s_n \,,$$

of (extended) features

$$s_i \subseteq V \,,$$

which are then called selection expressions, where each s_i selects a subset of the range $dom(a)$ of an attribute $a \in A$. A selection expression s is thus a Boolean function $E \rightarrow \{0, 1\}$ that is true if the value of the corresponding attribute is contained in the respective subset of V for the respective edge $e \in E$. The set of all selection expressions is denoted by S.
A *subgroup (extension)*

$$E_P := ext(P) := \{e \in E | P(e) = true\}$$

is the set of all edges which are covered by the pattern P. Using the set of edges, it is straightforward to extract the subset of covered nodes.

The interestingness of a pattern is determined by a quality function

$$q\colon 2^S \to \mathbb{R}\,.$$

It maps every pattern in the search space to a real number that reflects the interestingness of a pattern (or the extension of the pattern, respectively). Many quality functions for a single target feature, e. g., in the binary or numerical case, trade-off the size $n = |ext(P)|$ of a subgroup and the deviation $t_P - t_0$, where t_P is the average value of a given target feature in the subgroup identified by the pattern P and t_0 the average value of the target feature in the general population. Thus, standard quality functions are of the form

$$q_a(P) = n^a \cdot (t_P - t_0),\ a \in [0;1]\,.$$

For binary target concepts, this includes, for example, a *simplified binomial* function $q_a^{0.5}$ for $a = 0.5$, or the *gain* quality function q_a^0 with $a = 0$. However, as we will see below, such simple formalizations (as utilized by standard subgroup discovery approaches) do not cover the specific properties in dyadic network analysis - that is why provide specific adaptations for that case below.

While a quality function provides a *ranking* of the discovered subgroup patterns, often also a statistical assessment of the patterns is useful in data exploration. Quality functions that directly apply a statistical test, for example, the Chi-square quality function, e. g., [3] provide a p-value for simple interpretation.

For network data, there exist several quality measures for comparing a network structure to a null-model. For a given subgroup we can, for example, adapt common community quality measures, e. g., [7] for subgroup discovery. Also, the quadratic assignment procedure [32] (QAP) is a standard approach applying a graph correlation measure: For comparing two graphs G_1 and G_2, it estimates the correlation of the respective adjacency matrices M_1 and M_2 and tests that graph level statistic against a QAP null hypothesis [32]. QAP compares the observed graph correlation of (G_1, G_2) to the distribution of the respective resulting correlation scores obtained on repeated random row and column permutations of the adjacency matrix of G_2. However, this relates to the whole graph and not to specific subgroups of dyads, i. e., a subset of edges.

As we will see below, we can apply similar mechanisms for comparing a sub-network induced by a given subgroup pattern with a set of randomized sub-networks given the same distributional characteristics with respect to the total set of edges. However, in contrast to simple permutation operations, we have to take special care with respect to the social interaction properties, as we discuss below in detail, in order to compare the observed number of edges covered by a subgroup pattern with the expected number given a null-model.

Using a given subgroup discovery algorithm, the result of top-k subgroup discovery is the set of the k patterns $P_1, \ldots, P_k$, where $P_i \in 2^S$, with the highest interestingness according to the applied quality function. A subgroup discovery task can now be specified by the 5-tuple: (DB, c, S, q, k), where c indicates the target concept; the search space 2^S is defined by the set of basic patterns S.

4 Method

We first provide an overview on the proposed approach for the analysis of social interaction networks. Next, we present two novel quality functions for that task.

4.1 Compositional Network Analysis using Subgroup Discovery

We focus on the analysis of *social interaction networks* [2, 42], i. e., user-related social networks capturing social relations inherent in social interactions, social activities and other social phenomena which act as proxies for social user-relatedness. According to Wassermann and Faust [49, p. 37 ff.] social interaction networks focus on *interaction* relations between *people* as the corresponding actors. Then, a dyad, i. e., a link between two actors, models such a dyadic interaction. In a graph representation of the network, the dyad is then represented by an edge between two nodes (corresponding to the respective actors). Given attributed networks, also describing attributes, i. e., properties of nodes and/or edges can be used to characterize subgroups in order to *characterize* or *explain* a certain (observed) behavior, e. g., [21, 33, 49]. Here, we focus on *compositional network analysis* using subgroup discovery, where subgroups are induced by (a set of) describing attributes. Subgroup discovery enables hypotheses generation by directly exploring a given attribute space in order to identify interesting (compositional) subgroups according to some interestingness measure. As an exploratory method, we can e. g., focus on the top-k subgroups. Such patterns are then *local models* describing "interesting subsets" in terms of their attributes.

In the following, we focus on attributed networks, i. e., edge-attributed graphs with respect to actor attributes, enabling compositional dyadic analysis [49]. The interestingness can be flexibly defined using a quality measure. For social interaction networks, we distinguish between the following two properties:

1. Interaction duration: In social interaction networks, the duration of an interaction can be captured by a weight assigned to a specific link connecting the interacting actors. Then, simple networks that just capture those interactions can be represented by weighted graphs. In the unweighted case, we can just assign a default weight w for an edge e, e. g., $w(e) = 1.0$.
2. Interaction frequency: The frequency of interactions is typically indicated by multiple links between the two interacting actors, represented by a set of edges connecting the respective nodes in a multigraph. In addition, the duration of the interaction can also be captured as described above.

In the scope of this work, we focus on a numeric target feature t_P corresponding to the observed number of edges normalized by the expectation, for pattern P; for the interaction duration, we consider the weighted variant, i. e., taking the edge weights into acount. Then, we rank subgroups utilizing the (normalized) mean of that target feature t_P. It is important to note, that we use the number of all possible contacts (edges) for computing the mean of t_P, i. e., including edges with a zero weight. Therefore, we take into account all possible edges between all nodes (actors), as discussed below, for simple graphs (for interation duration), as well as for multigraphs where we also consider interaction frequency.

4.2 Quality Measures

For ranking a set of subgroup patterns, we propose two quality measures. Essentially, we distinguish two cases: First, simple compositional networks represented as simple attributed graphs, which can also be weighted, and second attributed multigraphs. We propose two quality functions for estimating *dyadic means* of a pattern P, corresponding to the numeric target feature t_P discussed above. This is combined with randomization approaches for estimating the significance of the respective values. Alltogether, this results in statistically well-founded quality functions, yielding intuitively interpretable values.

Simple Attributed Graphs In the case of a simple network (without multiple links) we can simply add up the number of (weighted) edges E_P captured by a pattern P, and normalize by the number of all possible edges n_E in the node subset induced by P, i. e., all contributing nodes that are connected by any edge e contained in E_P. That means, for example, that if we consider the mean duration of contacts in a social interaction network as the target t_P, where the duration is indicated by the weight of a (contact) edge between two nodes (i. e., the involved actors), then we normalize by the number of all possible contacts that can occur in that set of nodes. Thus, intuitively, we take contacts of length zero into account for completeness. Thus, for a pattern P, we estimate its quality $q_S(P)$ as follows:

$$q_S(P) = Z(\frac{1}{n_E} \cdot \sum_{e \in E_P} w(e)) \,, \tag{1}$$

with $n_E = \frac{n_{E_P}(n_{E_P}-1)}{2}$, where n_{E_P} is the number of nodes covered by a pattern P. Z is a function that estimates the statistical significance of the obtained value (i. e., t_P) given a randomized model, which we discuss below in more detail.

Attributed Multigraphs For more complex attributed networks containing multi-links between actors, we model these as attributed multigraphs. Then, we can additionally take the interaction frequency into account, as discussed above. The individual set of interactions is modeled using a set of links between the different nodes representing the respective actors of the network. Thus, for normalizing the mean of target t_P, we also need to take into account the multiplicity of edges between the individual nodes. Then, with $n_E = \frac{n_{E_P}(n_{E_P}-1)}{2}$ indicating the total number of (single) edges between the individual nodes captured by pattern P, $m_i, i = 1 \ldots n_E$ models the number of multi-edges for an individual edge i connecting two nodes. With that, extending Equation 1 for a pattern P in the multigraph case, we estimate its quality $q_M(P)$ as follows:

$$q_M(P) = Z(\frac{1}{n_E + m_E} \cdot \sum_{e \in E_P} w(e)) \,, \tag{2}$$

with $m_E = \sum_{i=1}^{n_E} (m_i - 1)$. It is easy to see that Equation 2 simplifies to Equation 1 for a simple attributed network, as a special case.

Randomization-Based Significance Estimation As summarized above in Section 3, standard quality functions for subgroup discovery compare the mean of a certain target concept with the mean estimated in the whole dataset. In the dyadic analysis that we tackle in this paper, however, we also need to take edge formation of dyadic structures into account, such that, e. g., simply calculating the mean of the observed edges normalized by all edges for the whole dataset is not sufficient. In addition, since we use subgroup discovery for identifying a dyadic subgraph (i. e., a set of edges) induced by a pattern, we also aim to confirm the *impact* by checking the statistical significance compared to a null-model. For that, we propose a sampling based procedure: We draw r samples without replacement with the same size of the respective subgroup in terms of the number of edges, i. e., we randomly select r subsets of edges of the whole graph. For the two cases discussed above, i. e., for the simple attributed graph and the multigraph representation, we distinguish two cases:

1. Simple graph network representation: In the simple case, we just take into account the
$$N = \frac{n(n-1)}{2}$$
possible edges between all nodes of the simple graph. Thus, in a sampling vector $R = (r_1, r_2, \ldots, r_N)$, we fill the $r_i, i = 1 \ldots N$ positions with the weights of the corresponding edges of the graph, for which that a non-existing edge in the given graph is assigned a weight of zero.
2. Multigraph network representation: In the multigraph case we also consider the number of all possible edges between all the nodes, however, we also need to take the multi-edges into account, as follows:
$$N = \frac{n(n-1)}{2} + \sum_{i=1}^{n} (m_i - 1)\,,$$
where $m_i, i = 1, \ldots, n$, are the respective multi-edge counts for an individual edge i. As above, we assign the sampling vector R accordingly, where we set the weight entries of non-existing edges to zero.

For selecting the random subsets, we apply sampling without replacement. This is essentially equivalent to a shuffling based procedure, e. g., [19,23]). Then, we determine the mean of the target feature t_R (e. g., mean duration) in those induced r subsets of edges. In that way, we build a distribution of "false discoveries" [19] using the r samples. Using the mean t_P in the original subgroup and the set of r sample means, we can construct a z-score which directly leads to statistical assessment for computing a p-Value. This is modeled using the function $Z(t_P), Z : \mathbb{R} \rightarrow \mathbb{R}$ which is then used for estimating the statistical significance of the target t_P of pattern P. In order to ensure that the r samples are approximately normally distributed, we can apply a normality test, for example, the Shapiro-Wilk-test [48]. If normality is rejected, a possible alternative is to compute the empirical p-value of a subgroup [23]. However, in practice often the distribution of the sampled means is approximately normally distributed, so that a p-value can be directly computed from the obtained z-score.

5 Results

Below, we describe the utilized two real-world datasets on social face-to-face interaction networks and experimental results of applying the presented approach.

5.1 Datasets

We applied social interaction networks captured at two scientific conferences, i. e., at the LWA 2010 conference in Kassel, Germany, and the Hypertext (HT) 2011 conference in Eindhoven, The Netherlands. Using the CONFERATOR system [5], we invited conference participants[1] to wear active RFID proximity tags. [2] When the tags are worn on the chest, tag-to-tag proximity is a proxy for a (close-range) face-to-face (F2F) contact, since the range of the signals is approximately 1.5 meters if not blocked by the human body, cf. [14] for details. We record a F2F contact when the length of a contact is at least 20 seconds. A contact ends when the proximity tags do not detect each other for more than 60 seconds. This results in time-resolved networks of F2F contacts. Table 1 provides summary statistics of the collected datasets; see [27] for a detailed description.

Table 1. Statistics/properties of the real-world datasets: Number of participants $|V|$, unique contacts $|U|$, total contacts $|C|$ average degree, diameter d, density, count of F2F contacts (C), cf. [27] for details.

Network	$\|V\|$	$\|U\|$	$\|C\|$	∅Degree	d	Density	$\|C\|$
LWA 2010	77	1004	5154	26.08	3	0.34	5154
HT 2011	69	550	1902	15.94	4	0.23	1902

In addition to the F2F contacts of the participants, we obtained further (socio-demographic) information from their Conferator online profile. In particular, we utilize information on the participants' (1) *gender*, (2) *country* of origin, (3) (university) *affiliation*, (4) academic status – *position* – i. e., professor, postdoc, PhD, student, (5) and their main conference *track* of interest. Note that not all attributes are available for both conferences (e. g., country is not available for the LWA 2010 conference since almost all participants were from Germany; here, we refer to the (university) affiliation instead. In contrast, the country information is very relevant for HT 2011. For those attributes given above, we created features on the edges of the attributed (multi-)graphs in such a way, so that an edge was labeled with "<feature>=EQ" if the respective nodes shared the same value of the feature, e. g., *gender=female* for both nodes. Otherwise, the edge was labeled with "<feature>=NEQ". That means that, for example, the subgroup described by the pattern *gender=EQ* contains the nodes, for which the dyadic actors always agree on their attribute *gender*.

[1] Study participants also gave their informed consent for the use of their data (including their profile) in scientific studies.

[2] http://www.sociopatterns.org

5.2 Experimental Results and Discussion

For compositional analysis, we applied subgroup discovery on the attributes described in Section 5.1. We utilized the VIKAMINE [9] data mining platform for subgroup discovery[3], utilizing the SD-Map* algorithm [8], where we supplied our novel quality functions for determining the top-20 subgroups.

For the target concept, we investigated the *mean length of contacts* – corresponding to the *duration* of a social interaction in the respective subgroup. We applied both simple attributed networks, and multigraph representations: For the former, social interactions between respective actors were aggregated, such that the corresponding weight is given by the sum of all interactions between those actors. For the multigraph case, we considered the face-to-face interations with their respective durations individually. Tables 2-5 show the results.

Table 2. Top-20 most exceptional subgroups according to the aggregated duration of face-to-face interactions at LWA 2010 (simple attributed network): The table shows the respective patterns, the covered number of dyads, the mean contact length in seconds and the significance compared to the null-model (Quality (Z)).

Description	Size	∅CLength	Quality (Z)
track=EQ	456	182.05	19.01
affiliation=NEQ	959	245.39	18.91
position=NEQ	885	227.44	17.93
affiliation=NEQ, position=NEQ	868	220.01	17.36
affiliation=NEQ, track=EQ	428	158.18	16.22
position=NEQ, track=EQ	392	145.7	15.71
gender=NEQ	705	182.5	15.43
affiliation=NEQ, position=NEQ, track=EQ	381	139.92	15.2
gender=NEQ, track=EQ	312	123.84	14.01
affiliation=NEQ, gender=NEQ	669	160.01	13.2
gender=NEQ, position=NEQ	627	152.02	12.89
affiliation=NEQ, gender=NEQ, position=NEQ	614	145	12.1
gender=EQ	299	257.69	11.91
gender=EQ, track=EQ	144	189.02	11.75
affiliation=NEQ, gender=NEQ, track=EQ	289	102.15	11.35
affiliation=NEQ, gender=EQ, track=EQ	139	179.23	11.25
affiliation=NEQ, gender=EQ, position=NEQ, track=EQ	120	179.59	11.13
gender=EQ, position=NEQ, track=EQ	123	180.46	11.06
affiliation=NEQ, gender=EQ	290	252.35	11.01
affiliation=EQ, track=EQ	28	298.74	11

Overall, we notice several common patterns in those tables, both for LWA 2010 and HT 2011: We observe the relatively strong influence of homophilic features such as *gender*, *track*, *country*, and *affiliation* in the detected patterns, confirming preliminary work that we presented in [11] only analyzing the individual features and their contribution to establishing social interactions. Using compositional subgroup discovery we can analyze those patterns at a more fine-grained level, also taking more complex patterns, i. e., combinations of different

[3] http://www.vikamine.org

features into account. Thus, our results indicate more detailed findings both concerning the individual durations, the influence of repeating interactions, and the impact of complex patterns given by a combination of several features.

Table 3. Top-20 most exceptional according to the non-aggregated duration of face-to-face interactions at LWA 2010 (attributed multigraph): The table shows the respective subgroup patterns, the covered number of dyads, the mean contact length in seconds and the significance compared to the null-model (Quality (Z)).

Description	Size	Length	Quality (Z)
affiliation=EQ, gender=EQ, position=EQ, track=EQ	30	239	793.96
affiliation=EQ, gender=EQ, position=NEQ, track=NEQ	7	71.29	491.59
affiliation=EQ, gender=EQ, position=EQ, track=NEQ	39	164.02	476.73
affiliation=EQ, gender=EQ, track=EQ	39	160.73	475.71
affiliation=EQ, gender=EQ, position=EQ	69	184.37	412.34
affiliation=EQ, gender=EQ, track=NEQ	46	127.68	341.41
affiliation=EQ, gender=NEQ, position=NEQ, track=NEQ	34	105.83	337.98
affiliation=EQ, gender=EQ, position=NEQ, track=EQ	9	44.63	274.97
affiliation=EQ, position=NEQ, track=NEQ	41	91.99	263.29
affiliation=EQ, gender=EQ	85	128.89	257.45
affiliation=EQ, position=EQ, track=NEQ	78	119.78	249.23
affiliation=EQ, gender=NEQ, position=EQ, track=NEQ	39	77.24	226.94
affiliation=EQ, gender=EQ, position=NEQ	16	44.93	203.45
affiliation=EQ, gender=NEQ, track=NEQ	73	86.25	182.48
affiliation=EQ, track=NEQ	119	103.35	171.08
affiliation=EQ, gender=NEQ, position=NEQ, track=EQ	98	92.89	170.31
gender=EQ, position=EQ, track=EQ	142	107.1	165.17
affiliation=NEQ, gender=EQ, position=EQ, track=NEQ	87	83.01	162.58
affiliation=EQ, gender=NEQ, position=EQ, track=EQ	228	135.41	161.12
affiliation=EQ, position=EQ, track=EQ	258	137.37	156.49

Furthermore, we also observe that the compositional multigraph analysis, i. e., focusing on dyadic interactions in the multigraph case focuses on much more specific patterns with many more contributing features, in contrast to more general patterns in the case of the simple attributed network. That is, for the multigraph case smaller subgroups (indicated by the size of the set of involved actors/nodes) are detected that are more specific regarding their descriptions, i. e., considering the length of the describing features. Then, these can provide more detailed insights into, e. g., homophilic processes. We can assess different specializations of competing properties, see e. g., lines #1 and #3 in Table 3. Also, the "specialization transition" between two patterns provides interesting insights, e. g., considering the patterns *affiliation=EQ, gender=EQ* (line #10) and *affiliation=EQ, gender=EQ, track=EQ* (line #4) shown in Table 3 which indicates the strong homophilic influence of the track feature. A similar pattern also emerges for HT 2011, regarding *country=EQ, gender=NEQ, position=EQ*; here both *track=NEQ* and *track=EQ* improve on the mean contact duration; the latter is considerably stronger, also in line with our expectations, e. g., cf. [11].

Table 4. Top-20 most exceptional subgroups according to the aggregated duration of face-to-face interactions at HT 2010 (simple attributed network): The table shows the respective patterns, the covered number of dyads, the mean contact length in seconds and the significance compared to the null-model (Quality (Z)).

Description	Size	Length	Quality (Z)
gender=EQ	357	114.76	15.76
gender=EQ, track=EQ	114	83.87	15.32
country=EQ, gender=EQ, track=EQ	35	111.75	14.21
country=EQ, track=EQ	42	89.74	13.89
track=EQ	185	70.4	13.73
country=EQ, gender=EQ, position=NEQ, track=EQ	18	140.52	12.98
country=EQ, gender=EQ	55	70.06	12.75
country=NEQ	470	87.76	12.61
country=EQ	80	56.51	12.59
position=NEQ	365	76.89	11.87
gender=EQ, position=EQ, track=EQ	46	68.43	11.8
country=EQ, position=NEQ, track=EQ	23	99.62	11.62
position=EQ	185	60.15	11.45
position=EQ, track=EQ	60	53.32	11.44
country=EQ, gender=EQ, position=NEQ	30	82.03	11.29
country=NEQ, gender=EQ	302	82.91	11.19
gender=EQ, position=EQ	136	61.91	10.81
gender=EQ, position=NEQ	221	71.43	10.52
gender=EQ, position=NEQ, track=EQ	68	58.42	10.13
track=NEQ	365	70.22	10.03
country=EQ, position=NEQ	50	45.89	9.86

Table 5. Top-20 most exceptional subgroups according to the non-aggregated duration of face-to-face interactions at HT 2011 (attributed multigraph): The table shows the respective subgroup patterns, the covered number of dyads, the mean contact length in seconds and the significance compared to the null-model (Quality (Z)).

Description	Size	Length	Quality (Z)
country=EQ, gender=NEQ, position=EQ, track=EQ	13	159.57	353.49
country=EQ, gender=NEQ, position=EQ, track=NEQ	32	126.3	173.93
country=EQ, gender=NEQ, position=EQ	45	102.51	120.37
country=EQ, gender=NEQ, position=NEQ, track=EQ	15	45.74	92.91
country=EQ, gender=EQ, position=EQ, track=NEQ	17	42.27	83.02
country=EQ, gender=NEQ, track=EQ	28	49.86	74.91
country=EQ, gender=EQ, position=EQ, track=EQ	113	85.67	65.45
country=EQ, position=EQ, track=EQ	126	85.04	62.09
country=EQ, position=EQ, track=NEQ	49	52.29	61.21
country=EQ, gender=EQ, position=EQ	130	59.27	45.2
country=NEQ, gender=NEQ, position=EQ, track=EQ	32	29.08	42.28
country=EQ, gender=EQ, position=NEQ, track=NEQ	38	31.69	41.84
gender=NEQ, position=EQ, track=EQ	45	30.63	38.17
country=EQ, gender=NEQ, track=NEQ	78	41.06	38.02
country=EQ, gender=EQ, position=NEQ, track=EQ	255	72.55	36.41
country=EQ, position=EQ	175	52.37	35.98
country=NEQ, gender=EQ, position=EQ, track=EQ	166	41.72	32.72
gender=EQ, position=EQ, track=EQ	279	52.69	32.33
country=EQ, gender=EQ, track=EQ	368	66.86	30.3
country=EQ, position=NEQ, track=EQ	270	60.25	30.29
position=EQ, track=EQ	324	43.21	27.79

6 Conclusions

In this paper, we formalized the problem of detecting compositional patterns in attributed networks, i. e., capturing dyadic subgroups that show an interesting behavior as estimated by a quality measure. We presented a novel approach adapting techniques of subgroup discovery and exceptional model mining [3, 7, 18]. Furthermore, we discussed estimation methods for ranking interesting patterns, and presented two novel quality measures for that purpose. Finally, we demonstrated the efficacy of the approach using two real-world datasets.

Our results indicate interesting findings according to common principles observed in social interaction networks, e. g., the influence of homophilic features on the interactions. Furthermore, the applied quality functions allow to focus on specific properties of interest according to the applied modeling method, e. g., whether a simple attributed network or a multigraph representation is applied. Furthermore, the proposed quality functions are statistically well-founded, and provide a statistical significance value directly, also easing their interpretation.

For future work, we aim to extend the concepts developed in this work towards multiplex networks, also taking into account temporal network dynamics. For that, we aim to consider methods for analyzing sequential patterns [4] as well as approaches for modeling and analyzing multiplex network approaches, e. g., [25, 47]. Finally, methods for testing specific hypothesis and Bayesian estimation techniques, e. g., [4, 13, 20] are further interesting directions to consider.

Acknowledgements

This work has been partially supported by the German Research Foundation (DFG) project “MODUS” under grant AT 88/4-1.

References

1. Agrawal, R., Srikant, R.: Fast Algorithms for Mining Association Rules. In: Proc. VLDB. pp. 487–499. Morgan Kaufmann (1994)
2. Atzmueller, M.: Data Mining on Social Interaction Networks. JDMDH **1** (2014)
3. Atzmueller, M.: Subgroup Discovery. WIREs DMKD **5**(1), 35–49 (2015)
4. Atzmueller, M.: Detecting Community Patterns Capturing Exceptional Link Trails. In: Proc. IEEE/ACM ASONAM. IEEE Press, Boston, MA, USA (2016)
5. Atzmueller, M., Benz, D., Doerfel, S., Hotho, A., Jäschke, R., Macek, B.E., Mitzlaff, F., Scholz, C., Stumme, G.: Enhancing Social Interactions at Conferences. it - Information Technology **53**(3), 101–107 (2011)
6. Atzmueller, M., Doerfel, S., Hotho, A., Mitzlaff, F., Stumme, G.: Face-to-Face Contacts at a Conference: Dynamics of Communities and Roles. In: Modeling and Mining Ubiquitous Social Media, LNAI, vol. 7472. Springer (2012)
7. Atzmueller, M., Doerfel, S., Mitzlaff, F.: Description-Oriented Community Detection using Exhaustive Subgroup Discovery. Inf. Sci. **329**(C), 965–984 (2016)
8. Atzmueller, M., Lemmerich, F.: Fast Subgroup Discovery for Continuous Target Concepts. In: Proc. ISMIS. LNCS, vol. 5722, pp. 1–15. Springer (2009)

9. Atzmueller, M., Lemmerich, F.: VIKAMINE - Open-Source Subgroup Discovery, Pattern Mining, and Analytics. In: Proc. ECML/PKDD. Springer (2012)
10. Atzmueller, M., Lemmerich, F.: Exploratory Pattern Mining on Social Media using Geo-References and Social Tagging Information. IJWS **2**(1/2), 80–112 (2013)
11. Atzmueller, M., Lemmerich, F.: Homophily at Academic Conferences. In: Proc. WWW 2018 (Companion). IW3C2 / ACM (2018)
12. Atzmueller, M., Mollenhauer, D., Schmidt, A.: Big Data Analytics Using Local Exceptionality Detection. In: Enterprise Big Data Engineering, Analytics, and Management. IGI Global, Hershey, PA, USA (2016)
13. Atzmueller, M., Schmidt, A., Kloepper, B., Arnu, D.: HypGraphs: An Approach for Analysis and Assessment of Graph-Based and Sequential Hypotheses. In: New Frontiers in Mining Complex Patterns. LNAI, vol. 10312. Springer (2017)
14. Barrat, A., Cattuto, C., Colizza, V., Pinton, J.F., den Broeck, W.V., Vespignani, A.: High Resolution Dynamical Mapping of Social Interactions with Active RFID. PLoS ONE **5**(7) (2010)
15. Bendimerad, A., Cazabet, R., Plantevit, M., Robardet, C.: Contextual Subgraph Discovery With Mobility Models. In: International Workshop on Complex Networks and their Applications. pp. 477–489. Springer (2017)
16. Bothorel, C., Cruz, J.D., Magnani, M., Micenkova, B.: Clustering Attributed Graphs: Models, Measures and Methods. Network Science **3**(03), 408–444 (2015)
17. Burt, R.S.: Cohesion Versus Structural Equivalence as a Basis for Network Subgroups. Sociological Methods & Research **7**(2), 189–212 (1978)
18. Duivesteijn, W., Feelders, A.J., Knobbe, A.: Exceptional Model Mining. Data Mining and Knowledge Discovery **30**(1), 47–98 (Jan 2016)
19. Duivesteijn, W., Knobbe, A.: Exploiting False Discoveries – Statistical Validation of Patterns and Quality Measures in Subgroup Discovery. In: Proc. ICDM. pp. 151–160. IEEE (2011)
20. Espín-Noboa, L., Lemmerich, F., Strohmaier, M., Singer, P.: JANUS: A Hypothesis-Driven Bayesian Approach for Understanding Edge Formation in Attributed Multigraphs. Applied Network Science **2**(1), 16 (2017)
21. Frank, O.: Composition and Structure of Social Networks. Mathématiques et sciences humaines. Mathematics and Social Sciences (137) (1997)
22. Geng, L., Hamilton, H.J.: Interestingness Measures for Data Mining: A Survey. ACM Computing Surveys **38**(3) (2006)
23. Gionis, A., Mannila, H., Mielikäinen, T., Tsaparas, P.: Assessing Data Mining Results via Swap Randomization. ACM Transactions on Knowledge Discovery from Data (TKDD) **1**(3), 14 (2007)
24. Günnemann, S., Färber, I., Boden, B., Seidl, T.: GAMer: A Synthesis of Subspace Clustering and Dense Subgraph Mining. In: KAIS. Springer (2013)
25. Kanawati, R.: Multiplex Network Mining: A Brief Survey. IEEE Intelligent Informatics Bulletin **16**(1), 24–27 (2015)
26. Kaytoue, M., Plantevit, M., Zimmermann, A., Bendimerad, A., Robardet, C.: Exceptional Contextual Subgraph Mining. Mach. Learning **106**(8), 1171–1211 (2017)
27. Kibanov, M., Atzmueller, M., Illig, J., Scholz, C., Barrat, A., Cattuto, C., Stumme, G.: Is Web Content a Good Proxy for Real-Life Interaction? A Case Study Considering Online and Offline Interactions of Computer Scientists. In: Proc. ASONAM. IEEE Press, Boston, MA, USA (2015)
28. Kibanov, M., Atzmueller, M., Scholz, C., Stumme, G.: Temporal Evolution of Contacts and Communities in Networks of Face-to-Face Human Interactions. Science China Information Sciences **57**(3), 1–17 (March 2014)

29. Klösgen, W.: Explora: A Multipattern and Multistrategy Discovery Assistant. In: Advances in Knowledge Discovery and Data Mining, pp. 249–271. AAAI (1996)
30. Klösgen, W.: Applications and Research Problems of Subgroup Mining. In: Proc. ISMIS. pp. 1–15. Springer (1999)
31. Klösgen, W.: Handbook of Data Mining and Knowledge Discovery, chap. 16.3: Subgroup Discovery. Oxford University Press, New York (2002)
32. Krackhardt, D.: QAP Partialling as a Test of Spuriousness. Social Networks **9**, 171–186 (1987)
33. Lau, D.C., Murnighan, J.K.: Demographic Diversity and Faultlines: The Compositional Dynamics of Organizational Groups. Academy of Management Review **23**(2), 325–340 (1998)
34. Leman, D., Feelders, A., Knobbe, A.: Exceptional Model Mining. In: Proc. ECML/PKDD. LNCS, vol. 5212, pp. 1–16. Springer (2008)
35. Lemmerich, F., Atzmueller, M., Puppe, F.: Fast Exhaustive Subgroup Discovery with Numerical Target Concepts. Data Mining and Knowledge Discovery **30**, 711–762 (2016). https://doi.org/10.1007/s10618-015-0436-8
36. Lemmerich, F., Becker, M., Atzmueller, M.: Generic Pattern Trees for Exhaustive Exceptional Model Mining. In: Proc. ECML/PKDD 2012. Springer (2012)
37. Lemmerich, F., Becker, M., Singer, P., Helic, D., Hotho, A., Strohmaier, M.: Mining Subgroups with Exceptional Transition Behavior. In: Proc. ACM SIGKDD. pp. 965–974. ACM (2016)
38. Macek, B.E., Scholz, C., Atzmueller, M., Stumme, G.: Anatomy of a Conference. In: Proc. ACM Hypertext. pp. 245–254. ACM (2012)
39. McPherson, M., Smith-Lovin, L., Cook, J.M.: Birds of a Feather: Homophily in Social Networks. Annual Review of Sociology **27**(1), 415–444 (2001)
40. Mitzlaff, F., Atzmueller, M., Benz, D., Hotho, A., Stumme, G.: Community Assessment using Evidence Networks. In: Analysis of Social Media and Ubiquitous Data. LNAI, vol. 6904. Springer (2011)
41. Mitzlaff, F., Atzmueller, M., Hotho, A., Stumme, G.: The Social Distributional Hypothesis. Journal of Social Network Analysis and Mining **4**(216), 1–14 (2014)
42. Mitzlaff, F., Atzmueller, M., Stumme, G., Hotho, A.: Semantics of User Interaction in Social Media. In: Complex Networks IV, SCI, vol. 476. Springer (2013)
43. Morik, K.: Detecting Interesting Instances. In: Hand, D., Adams, N., Bolton, R. (eds.) Pattern Detection and Discovery, LNCS, vol. 2447, pp. 13–23 (2002)
44. Moser, F., Colak, R., Rafiey, A., Ester, M.: Mining Cohesive Patterns from Graphs with Feature Vectors. In: SDM. vol. 9, pp. 593–604. SIAM (2009)
45. Neely, R., Cleghern, Z., Talbert, D.A.: Using Subgroup Discovery Metrics to Mine Interesting Subgraphs. In: Proc. FLAIRS. pp. 444–447. AAAI (2015)
46. Robins, G., Pattison, P., Kalish, Y., Lusher, D.: An Introduction to Exponential Random Graph (p*) Models for Social Networks. Social Networks **29**(2) (2007)
47. Scholz, C., Atzmueller, M., Barrat, A., Cattuto, C., Stumme, G.: New Insights and Methods For Predicting Face-To-Face Contacts. In: Proc. ICWSM. AAAI (2013)
48. Shapiro, S.S., Wilk, M.B.: An Analysis of Variance Test for Normality (complete samples). Biometrika **52**(3/4), 591–611 (1965)
49. Wasserman, S., Faust, K.: Social Network Analysis: Methods and Applications. No. 8 in Structural Analysis in the Social Sciences, CUP, 1 edn. (1994)
50. Wrobel, S.: An Algorithm for Multi-Relational Discovery of Subgroups. In: Proc. PKDD. pp. 78–87. Springer (1997)
51. Wrobel, S., Morik, K., Joachims, T.: Maschinelles Lernen und Data Mining. Handbuch der Künstlichen Intelligenz **3**, 517–597 (2000)