ANALYSIS OF THE EVOLUTION OF RESEARCH GROUPS AND TOPICS IN THE ISMIR CONFERENCE

Mohamed Sordo
Center for Computational Science
University of Miami
msordo@miami.edu

Mitsunori Ogihara
Dept. of Computer Science
University of Miami
ogihara@cs.miami.edu

Stefan Wuchty
Dept. of Computer Science
University of Miami
wuchtys@cs.miami.edu

ABSTRACT

We present an analysis of the topics and research groups that participated in the ISMIR conference over the last 15 years, based on its proceedings. While we first investigate the topological changes of the co-authorship network as well as topics over time, we also identify groups of researchers, allowing us to investigate their evolution and topic dependence. Notably, we find that large groups last longer if they actively alter their membership. Furthermore, such groups tend to cover a wider selection of topics, suggesting that a change of members as well as of research topics increases their adaptability. In turn, smaller groups show the opposite behavior, persisting longer if their membership is altered minimally and focus on a smaller set of topics. Finally, by analyzing the effect of group size and lifespan on research impact, we observed that papers penned by medium sized and long lasting groups tend to have a citation advantage.

1. INTRODUCTION

Music Information Retrieval (MIR) is an interdisciplinary research field that integrates a wide variety of research areas, including audio signal processing, musicology, music psychology and cognition, information retrieval, and human-computer interfaces. The collection of papers published in the annual proceedings of the ISMIR conference provides a wealth of information enabling us to mine for knowledge such as the networks of researchers that contribute papers and corresponding topics. Specifically, such abundant data allows us to explore two main research questions. First we focus on topics in the field. Given the breadth of the expertise of the field and the high speed at which the digital technologies are developing, we investigate if popular topics can be transient. Second, we study the stability of research groups that emerge from the co-authorships of manuscripts, focusing on their sizes, diversity of topics and competitiveness. While various approaches for the exploration of knowledge in the ISMIR paper collection exist, we considered a combination of network and text analysis.

Recently, the analysis of scientific endeavors by investigating author relationships and their manuscripts provided insights into innovation and idea creation processes [3], inter-dependencies between disciplines [2], or potential high-impact discoveries [5]. Utilizing proceedings of the mainstream conference we provide such a map of the status and temporal evolution of the MIR field. To the best of our knowledge, only two studies on the nature of the ISMIR proceedings [7,10] have been presented recently. Grachten et al. [7] applied text mining techniques and non-negative matrix factorization to identify topics and study their evolution over time. Lee et al. [10] applied simple text statistics to detect topics in paper titles and abstracts. In our case, we present a much broader analysis of research topics, that we map to categories that were defined by the ISMIR community. While Lee et al. [10] also presented a few statistics to identify patterns of co-authorship we model the co-authorship as a complex network and study its topology. Yet, the main contribution of this paper is the identification of research groups and their evolution over time, and especially their time and topic dependencies. In the context of this paper, we do not use the definition of a research group in its traditional sense (e.g., a research institution). Rather, we define it as a topological group in a co-authorship network.

As for the organization of the paper we first describe the network of collaborations among authors in section 2. In section 3, we provide an analysis of the manuscripts text contents with a generative mixture model, allowing us to find temporal trends in the popularity of topics over the years. In section 4 we identify research groups in the co-authorship network and analyze their evolution throughout the lifetime of the conference, investigating their time and topic dependencies. Finally, we discuss our findings in section 5.

2. CO-AUTHORSHIP ANALYSIS

Utilizing all manuscripts in the proceedings of the ISMIR conference from 2000-2014 we observed that the mean number of authors per manuscript is growing over time (Table 1), confirming previous results [18]. Starting with the proceedings of the 2000 conference, we added new manuscripts that were published in a given year to a grow-

Year	**Papers**	**Authors**	**Authors/Paper**
2000	35	63	1.94
2001	37	82	2.54
2002	53	113	2.36
2003	47	108	2.74
2004	104	213	2.41
2005	114	232	2.73
2006	87	185	2.56
2007	127	267	2.84
2008	105	262	2.93
2009	123	292	3.05
2010	110	262	3.01
2011	133	320	2.97
2012	101	264	3.21
2013	98	232	3.02
2014	106	273	3.24

Table 1. For each year, we show the total number of papers and authors that published a manuscript in the proceedings of the ISMIR conference.

Year	**N**	$\langle \mathbf{k} \rangle$	**C**	$\langle \mathbf{d} \rangle$	**SGC**	**D**
2000	63	1.81	0.47	1.00	9.52%	1
2001	129	2.51	0.55	1.00	6.20%	1
2002	202	2.62	0.55	3.20	10.40%	6
2003	268	2.86	0.55	3.22	8.21%	6
2004	400	2.92	0.58	4.14	10.75%	10
2005	522	3.18	0.59	3.96	14.75%	9
2006	625	3.18	0.60	4.34	14.72%	10
2007	756	3.34	0.62	4.85	20.24%	11
2008	884	3.44	0.64	7.72	41.18%	17
2009	1041	3.58	0.65	8.13	46.11%	18
2010	1170	3.70	0.66	6.60	48.55%	15
2011	1339	3.76	0.67	6.47	53.70%	15
2012	1442	3.94	0.68	5.82	58.46%	14
2013	1548	4.03	0.69	5.74	61.18%	13
2014	1683	4.14	0.70	5.52	60.90%	13

Table 2. We show properties of the cumulative authors' collaboration networks, combining manuscripts up to a given year. In particular, N is the number of nodes, $\langle k \rangle$ is the mean degree, C is the clustering coefficient. Furthermore, $\langle d \rangle$ is the avg. shortest path of the SGC, which stands for the size (percentage of nodes) of the strong giant component, while D is the diameter of the SGC.

ing pool of papers. Based on such cumulative sets of manuscripts, we constructed undirected unweighted networks G, where nodes represent authors, while edges indicate their co-authorships up to a given year.

Table 2 suggests that the cumulative networks drastically increased in size over time, a statistics that coincides with an increasing number of collaboration partners (i.e. mean degree $\langle k \rangle$).

Another important measure of social networks is the clustering coefficient, reflecting the transitivity of a network. In particular, this network parameter determines the fraction of edges that appear between the neighbors of a given author over all such possible links [17]. Table 2 indicates that the co-authorship networks appear increasingly clustered, resembling a well known feature of other social networks from different domains [9, 12]. Such a high level of clustering may be rooted in the assumption that many authors work in the same research field, and as a consequence, are aware of each others work [13]. Another possible explanation may be that authors tend to write papers with colleagues from the same institution. Furthermore, we stress that our way of constructing a network of collaborations between authors emphasizes manuscripts with a large number of authors. Specifically, a set of authors that penned a manuscript together is represented as a clique, a graph that has a clustering coefficient of 1. Consequently, manuscripts with many authors potentially introduce a bias toward strongly clustered networks.

Another network parameter that well reflects the underlying topology of an emerging network over time is the Strong Giant Component, SGC, defined as the greatest connected subset of nodes in a network. In particular, a high value of SGC points to the observation that the vast majority of scientists are connected through mutual collaborations. During the first years of the conference (up until 2007), Table 2 indicates that the size of SGCs was small, suggesting that collaborations between authors appeared rather scattered. However, the size of the SGC doubled in 2008, indicating an increased convergence where previously present authors increasingly published a manuscript together. On the other hand, the observed increase in size also points to a gradual increase in the mean shortest path $\langle d \rangle$ between all pairs of nodes in the SGC. A closer look at our data confirmed that the increase in size of the SGC was the consequence of a merger of the two largest components from the previous year. Notably, this topological change was caused by a small set of nodes that bridged the previously disconnected components in the underlying network. As a consequence, the topological mean shortest path lengths between nodes increased substantially since shortest paths between nodes that were placed in previously disjoint components run through the small set of connecting nodes. Such an assumption is further confirmed by the increasing diameter of the underlying networks defined as the maximum of shortest paths through a given network (Table 2).

3. RESEARCH TOPICS

The analysis of the time evolution of research topics is a valuable asset for a research community to solve initial problems and to adapt to challenging areas of research. In this section, we automatically extract underlying topics from the text content of proceeding papers, allowing us to map the evolution of these topics since the inception of the MIR field.

topic	**00**	**01**	**02**	**03**	**04**	**05**	**06**	**07**	**08**	**09**	**10**	**11**	**12**	**13**	**14**
MIR Data & fundamentals															
mus. signal processing	17.1	-	-	-	-	-	8.0	-	-	**19.5**	-	-	10.9	10.2	12.3
metadata, & semantic web	11.4	5.6	-	**17.0**	12.5	11.5	**16.1**	12.7	10.5	9.8	11.8	9.8	-	-	-
social tags & user gen. data	-	-	-	-	-	-	-	**13.5**	10.5	12.2	10.9	**12.8**	11.9	**12.2**	-
lyrics & genres & moods	-	-	-	-	-	-	-	-	11.4	11.4	-	10.5	9.9	-	11.3
Domain Knowledge															
comp. music. & ethnomus.	-	8.3	-	-	-	-	-	-	-	-	-	-	-	-	-
mus. notation	-	8.3	-	-	-	-	-	-	-	-	-	-	-	-	-
mir & cultures	-	-	-	-	-	-	-	-	-	-	-	9.8	-	10.2	-
Mus. Features & Properties															
melody & motives	11.4	-	11.3	8.5	8.7	-	9.2	11.9	-	-	-	11.3	12.9	-	-
harmony, chords & tonality	-	13.9	-	-	**13.5**	8.8	9.2	10.3	9.5	13.0	10.9	10.5	11.9	10.2	-
rhythm, beat, tempo	-	**19.4**	-	12.8	**13.5**	12.4	-	-	**13.3**	8.9	11.8	-	-	**12.2**	12.3
mus. affect, emot. & mood	-	-	-	10.6	-	-	-	-	-	-	-	-	-	10.2	-
structure, segment. & form	-	-	11.3	-	-	-	-	-	10.5	12.2	10.0	12.0	8.9	10.2	13.2
Music Processing															
sound source separation	-	-	-	-	-	-	8.0	10.3	-	-	**13.6**	-	**14.9**	**12.2**	11.3
mus. transcrip. & annot.	5.7	8.3	-	-	-	11.5	-	-	-	-	-	-	-	**12.2**	-
optical mus. recognition	-	-	-	-	-	-	6.9	10.3	-	-	-	-	9.9	-	-
align., synch. & score foll.	-	-	-	10.6	-	12.4	-	-	-	-	-	-	-	-	-
mus. summarization	-	-	7.5	-	-	-	-	-	-	-	-	-	-	-	-
fingerprinting	-	-	11.3	-	-	-	-	-	-	-	-	**12.8**	-	-	-
automatic classification	8.6	11.1	11.3	12.8	**13.5**	**14.2**	13.8	12.7	12.4	13.0	11.8	-	-	-	**14.2**
indexing & querying	**22.9**	13.9	9.4	10.6	7.7	9.7	9.2	-	-	-	10.9	-	-	-	-
pattern match. & detection	-	11.1	-	8.5	10.6	9.7	-	-	11.4	-	-	-	-	-	5.7
similarity metrics	-	-	-	8.5	9.6	-	11.5	8.7	-	-	8.2	10.5	-	-	-
Application															
user behavior & modeling	-	-	-	-	-	-	-	-	-	-	-	-	8.9	-	-
digital libraries & archives	11.4	-	-	-	10.6	-	-	-	-	-	-	-	-	-	-
mus. retrieval systems	-	-	**22.6**	-	-	-	-	-	10.5	-	-	-	-	-	8.5
mus. rec. & playlist gen.	-	-	15.1	-	-	9.7	8.0	-	-	-	-	-	-	-	11.3
mus. & gaming	-	-	-	-	-	-	-	9.5	-	-	-	-	-	-	-
mus. software	11.4	-	-	-	-	-	-	-	-	-	-	-	-	-	-

Table 3. Utilizing a LDA model, we determined topic evolution over time, where topics are grouped according to the topic classification in the call for papers. Values in bold correspond to the most salient topics in each ISMIR conference edition.

3.1 Topic extraction

We automatically extract the main topics by using Latent Dirichlet Allocation (LDA) [4], a generative probabilistic model in which documents are represented as random mixtures over latent topics. Each topic is characterized by a multinomial distribution over words that form those documents [4]. As a main characteristic LDA assumes that the topic distribution has a Dirichlet prior, which not only results in a smooth distribution but also simplifies the problem of topic inference [16].

In particular, we used the MALLET implementation of LDA, a java-based package for statistical natural language processing, document classification, clustering, topic modeling and machine learning applications of text [11]. MALLET's implementation takes a text corpus and the number of topics (k) to generate as input, and produces a list of the most relevant topics for that corpus, along with the topics' most salient terms. Furthermore, MALLET also provides a distribution of the topics among the documents that form the corpus and includes a text pre-processing step prior to generate the topic models.

Here, we build a corpus for each set of manuscripts in the ISMIR proceedings in a given year and set $k = 10$, resembling the number of oral sessions defined by the program chairs, which typically group paper presentations by their topic affinity. For the text pre-processing step, we removed English stopwords, considered words that were longer than 2 characters and used a combination of word unigrams and bigrams. Since topics produced by an LDA model are only described by their word distribution, we manually assigned "titles" after an inspection of the most probable terms. In particular, we used the list of topics described in the conference call for papers [1] as our basis to assign and disambiguate topic titles [2]. We also observed that this LDA implementation was systematically producing a topic containing most of the common words in any MIR publication (such as *music*, *system*, *information*, *query*, *retrieval*). Since such topics were almost never the most salient topic of a document in the corpus we removed them from our analysis.

3.2 Topic evolution

Table 3 shows the most salient topics that appeared in the ISMIR proceedings over time, as well as a visualization of their evolution, pointing to their presence in each conference edition. Each value in Table 3 represents the percentage of papers per year whose most probable topic in the

[1] `http://ismir2015.uma.es/callforpapers.html`

[2] due to lack of space we made the topic distribution available online: `https://goo.gl/6OmGl5`

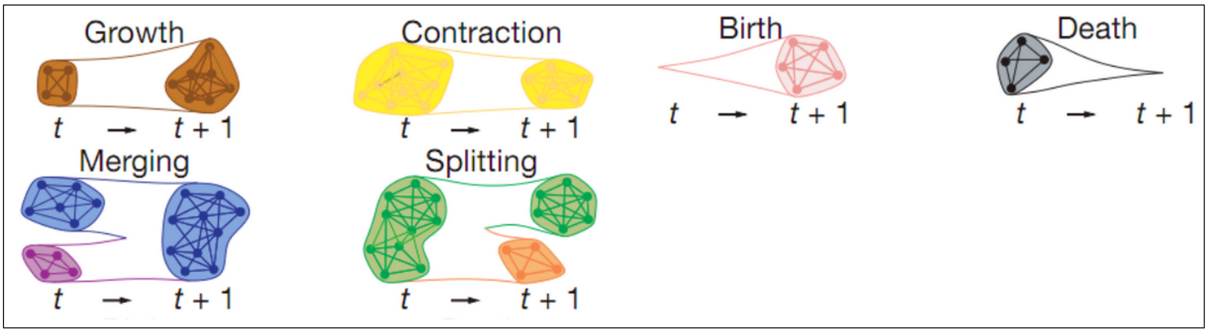

Figure 1. Fission/fusion patterns in social networks from [14]. Considering social networks over time, groups are governed by dynamic events such as mergers (i.e .fusion) and splits of groups (fission).

document–topic distribution corresponds to the topic in a given row. For instance, the topic *indexing and querying* was the most salient topic of 22.9% of the papers in the ISMIR 2000 edition. We stress that the lack of a value for a topic in a specific conference edition does not necessarily point to its absence in the underlying edition. In fact, such an observation rather indicates that the topic in question was not among the $k = 10$ most salient topics that year.

For a better interpretation, we grouped topics according to the topic classification in the call for papers. Notably, we observed that the most salient topics over time belonged to the categories “MIR Data and Fundamentals”, “Musical Features and Properties” and “Music Processing”, respectively, categories that can be regarded as the core categories in the MIR field. Some topics have been largely present over time, such as automatic classification, harmony, melody, etc. Other topics appeared or became more popular halfway through the life-span of the conference (e.g. tags, lyrics, moods, structure, etc.) or in the last few years (e.g. source separation and music cultures). Such observations may be the consequence of introducing emerging research topics or approaches from “neighboring” communities or from a shift in research funding by national or international agencies. Finally, some topics emerged that have only been present in a short time period (e.g. digital libraries or music and gaming). In particular, we highlight the *digital libraries* topic, which was more present during the first editions of the conference, but disappeared from the most salient topics over time. Such an observation may be explained by the increased focus on music content- and context-based analysis (groups 1, 3 and 4).

4. GROUP DETECTION AND EVOLUTION

In the past few years, considerable attention has been paid to uncovering topological groups in social networks. These groups are expected to fundamentally impact the network’s dynamical properties as well: nodes that belong to the same tightly connected module are expected to display highly correlated dynamical activity, compared to nodes belonging to different groups. Previous studies found that large groups are more stable and have a longer lifetime if they are capable of dynamically altering their membership, suggesting that an ability to change the group composition results in better adaptability [14]. Small groups display the opposite trend, suggesting that their condition for stability is an unchanged group composition. These discoveries are expected to play a fundamental role in our understanding of human dynamics, with particular impact on our ability to detect persuasion campaigns in a changing network environment. Notably, dynamics of group composition have been noted by Dunbar and co-workers as a key mechanism to understand underlying human behavior across domains [1, 6]. In particular, we not only expect that such patterns will occur in the co-authorship networks based on conference proceedings of the ISMIR conference but also assume that the (in)stability of groups is a function of their underlying topics.

4.1 Method

Our method is a modification of the method presented in [14]. In particular, we define a co-authorship network for each edition of the conference, where each edge represents a manuscript that a pair of authors penned in a given year. Furthermore, we extract groups using the clique percolation method (CPM), an algorithm for the detection of overlapping network communities [15]. Groups in CPM, called k-clique percolation clusters, are built up from adjacent k-cliques[3]. Two k-cliques are considered adjacent if they share $k-1$ nodes. Such a definition allows nodes to appear in several k-clique percolation clusters, a suitable assumption, given that authors may participate in more than one group. Specifically, we set $k = 3$, since papers in the ISMIR proceedings are co-authored on average by 3 scientists. As a consequence, this restriction implies that authors who collaborate with less than 2 other authors will never be part of a group.

After groups have been determined in a given year, we need to find their possible matches in subsequent years. In particular, we construct a joint network by merging nodes and edges of networks at consecutive time steps t and $t+1$ [14], considering different fission/fusion patterns (Fig. 1). We label the set of groups in time t as D, the set of groups in time $t+1$ as E, and the set of groups in the joint network

[3] subgraphs of size k in which each node is connected to every other nodes

life span	1	2	3	4	5	6	8	9
num. groups	327	65	24	7	2	5	1	1

Table 4. Distribution of groups by their life time

as V. The definition of CPM implies that each group in D (E) is contained in exactly one group in V, although not all groups in V will contain a group in D (E). If a group $V_k \in V$ contains a group $E_j \in E$ but no group in D, then group E_j is considered born. Similarly, if a group $V_k \in V$ contains a group $D_i \in D$ but no group in E, then group D_i is considered dead. Furthermore, if a group $V_k \in V$ contains one or more groups in D and one or more groups in E, then the relative overlap between all different pairs (D^k_i, E^k_j) is obtained as:

$$C^k_{i,j} = \frac{D^k_i \cap E^k_j}{D^k_i \cup E^k_j} \tag{1}$$

The pair (D^k_i, E^k_j) of groups that maximizes this formula is considered a match of the same group in consecutive time steps t and $t+1$. The remaining groups are either marked as dead (D^k) or born (E^k).

Contrary to the approach in [14], we considered the overlap of nodes (instead of edges) to check whether groups in D and E are contained in V. Since our networks are built at discrete times two networks at time steps t and $t+1$ do not necessarily have a high overlap as suggested in [14]. Although the overlap of nodes might incur more noisy matchings [14], we observed that our approach is less sensitive to the noise since the overlap of networks is limited.

In some cases, we may consider a group dead in time step $t+1$ but observe it re-born in time step $t+2$ as a consequence of members that did not publish a paper in the proceedings of year $t+1$. Even though the time interval is larger than one year we consider them as the same group. Specifically, we matched such dead groups at time t with born groups at $t+n$ with $n = 2$, as larger values of n only merged a very small number of groups.

4.2 Experimental results

Applying our method to our set of ISMIR proceedings we obtained a list of 432 groups, distributed as shown in table 4. Notably, we observed that only 40 groups persisted for 3 or more years, representing less than 10% of the total. In particular, we split the groups in two categories: groups with short (< 3) and a long life spans (≥ 3). Analyzing group sizes we determined the average size of each group (in terms of group members in each time step) over time and split the groups in three size categories: small (avg. size < 4 members), medium ($4 \leq$ avg. size < 5) and large (avg. size > 5). The three size categories contained 234, 120 and 78 groups, respectively.

Group size	$\sigma_{\mathbf{avg}}$	**Avg. cumul. authors**
small	0.58	5.17±1.47
medium	1.18	9.06±2.54
large	2.35	13.55±3.47

Table 5. Group member variability in groups with long life time.

Group size	μ (σ)	**median**
small	3.17±1.07	3
medium	3.88±1.60	3
large	6.0±2.04	5

Table 6. Topic variability in groups with long life time.

4.2.1 Group member variability

We analyzed the variability of group members when groups persisted for a longer period of time. In particular, we calculated the variance of group size for each group in each group category, and averaged them using $\sigma_{avg} = \sqrt{\sum_t var(s_t)}$, where s_t is the size of a group at a particular time t. Moreover, we computed the average number of distinct authors that participated in the group at a given time. Table 5 indicates that larger groups tend to have a higher variability of members to persist for a longer period of time. Notably, we observed the opposite when we considered small groups, confirming results in [14].

4.2.2 Topic variability

A higher topic variability means that groups change topics constantly throughout their life time. In particular, we calculated the average number of topics covered by small, medium and large groups (Table 6). Similar to the previous experiment, we only considered groups with a life time ≥ 3. To persist longer, large groups tend to cover more topics as exemplified by a higher topic variability, as opposed to medium or small groups. Such observations suggest that the persistence of groups does not only depend on their member dynamics, but also on the variability of research topics.

4.2.3 Group characteristics and scientific impact

Focusing on the relation between group characteristics and scientific impact we considered the number of citations of each paper, as of Google Scholar, representing an indicator of scientific impact. Specifically, we group papers by their most salient topic and only select the top 10 most cited papers in each topic, providing a total of 243 papers from 28 different topics [4] . Out of this set of 243 papers, we observed that only 137 were published by groups while the remainder was penned by one or two authors. As presented in Table 7 we observed that papers written by medium sized groups tend to get significantly more citations than

[4] some topics are present in less than 10 papers.

Group size/lifespan	avg. paper citations	# papers
small	62.54±54.74	54
medium	113.67±107.56	37
large	64.65±40.71	46
short	73.49±67.37	95
long	85.12±83.77	42

Table 7. Relation between group characteristics (size and life span) and scientific impact. We only consider the top 10 most cited papers per topic.

other group categories. As for aspects of a group's life time, papers by groups that last longer tend to get more citations than short living groups. Such an observation may be rooted in the assumption that persisting research groups with stable members may have a higher chance of getting noticed by their peers, positively affecting their research impact. Furthermore, we stress that the distribution of citations has heavy tails [18]. As a consequence the number of citations of highly cited papers varies widely, explaining the large margin of error in our analysis.

5. CONCLUSIONS

In this paper, we analyzed the evolution of the MIR field represented by the proceedings of its most prestigious conference ISMIR over the last 15 years. Notably, we found that the co-authorship network indicated a converging field of authors as indicated by the emergence of large connected and clustered network components as well as a trend toward larger research teams. While such a trend may be rooted in the way we constructed the network of co-authorships, our results also suggest that authors that have previously published conference papers separately increasingly collaborate. Therefore, present conference contributions may be viewed as 'seeds' for future collaborations between researchers that have not yet worked together. Assuming that increasing levels of collaboration govern innovation and the development of a research field, our results indicate that the ISMIR conference is a potential driver of the Music Information Retrieval field.

Furthermore, a topic analysis revealed persistent as well as 'rising' and 'falling' research topics over the years, providing a simple assessment of ISMIR's evolution. Such an analysis allowed us to investigate the longevity as well as the salience of certain topics. Our results also indicate the emergence of novel topics that potentially may dominate the focus of conference contributions in the future. Moreover, we assumed that the evolution of topics may be a function of the underlying groups of co-authors, prompting us to analyze their composition. Notably, we found that large groups persist through higher variability of team members while small groups show the opposite behavior. Furthermore, large groups show more variability of topics as opposed to medium or small groups. While not necessarily a function of group size, such results suggest that the variability of group composition may be the driving factor of topic variability. In particular, such results support the notion that groups composed of incumbents and newcomers have a heightened chance of success [8]. As a consequence, our results suggest that large transient groups may be the drivers for innovation given that such groups provide topic variability. In turn, the arrival of new members of a group may be accompanied by the introduction of new topics. As such, our observations also suggest that group persistence is not only a question of the variability of team members but also of research topics, ultimately providing a competitive edge.

6. ACKNOWLEDGMENTS

Mohamed Sordo acknowledges Òscar Celma, Amélie Anglade, Perfecto Herrera and Simon Dixon, with whom he collaborated a few years ago on an unpublished manuscript that influenced the first part of this paper.

7. REFERENCES

[1] Filippo Aureli, Colleen M. Schaffner, Christophe Boesch, Simon K. Bearder, Josep Call, Colin A. Chapman, Richard Connor, Anthony Di Fiore, Robin I. M. Dunbar, and S. Peter et al. Henzi. Fission-fusion dynamics. *Current Anthropology*, 49(4):627–654, 2008.

[2] Leana Bellanca. Measuring interdisciplinary research: analysis of co-authorship for research staff at the University of York. *Bioscience Horizons*, 2(2):99–112, 2009.

[3] Luis M. A. Bettencourt, David I. Kaiser, and Jasleen Kaur. Scientific discovery and topological transitions in collaboration networks. *Journal of Biometrics*, 3(3):210–221, 2009.

[4] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022, 2003.

[5] Chaomei Chen, Yue Chen, Mark Horowitz, Haiyan Hou, Zeyuan Liu, and Don Pellegrino. Towards an explanatory and computational theory of scientific discovery. *Journal of Informetrics*, 3(3):191–209, 2009.

[6] Robin I. M. Dunbar. Social cognition on the internet: testing constraints on social network size. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1599):2192–2201, 2012.

[7] Maarten Grachten, Markus Schedl, Tim Pohle, and Gerhard Widmer. The ismir cloud: A decade of ismir conferences at your fingertips. In *10th International Society for Music Information Retrieval Conference*, pages 63–68, 2009.

[8] Roger Guimera, Brian Uzzi, Jarret Spiro, and Luis Amaral. Team assembly mechanisms determine collaboration network structure and team performance. *Science*, 308(5722):697–702, 2005.

[9] Jon M. Kleinberg. Navigation in a small world. *Nature*, 406:845, 2000.

[10] Jin Ha Lee, M. Cameron Jones, and J. Stephen Downie. An analysis of ismir proceedings: Patterns of authorship, topic, and citation. In *10th International Society for Music Information Retrieval Conference*, pages 57–62, 2009.

[11] Andrew K. McCallum. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu, 2002.

[12] Mark E. J. Newman. The structure and function of complex networks. *SIAM Review*, 45(2):167–256, 2003.

[13] Mark E. J. Newman. Who is the best connected scientist? A study of scientific coauthorship networks. *Lecture Note in Physics-New York then Berlin-*, 650:337–370, 2004.

[14] Gergely Palla, Albert-László Barabási, and Tamás Vicsek. Quantifying social group evolution. *Nature*, 446(7136):664–667, 2007.

[15] Gergely Palla, Imre Derényi, Illés Farkas, and Tamás Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435(7043):814–818, 2005.

[16] Mark Steyvers and Tom Griffiths. Probabilistic topic models. *Handbook of latent semantic analysis*, 427(7):424–440, 2007.

[17] Duncan J. Watts and Steven H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442, June 1998.

[18] Stefan Wuchty, Benjamin F. Jones, and Brian Uzzi. The increasing dominance of teams in production of knowledge. *Science*, 316(5827):1036–1039, 2007.