Understanding Multilingual Social Networks in Online Immigrant Communities

Evangelos E. Papalexakis
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, USA
epapalex@cs.cmu.edu

A. Seza Doğruöz
Tilburg University
P.O. Box 90153
Tilburg, The Netherlands
a.s.dogruoz@gmail.com

ABSTRACT

There are more multilingual speakers in the world than monolingual ones. Immigration is one of the key factors to bring speakers of different languages in contact with each other. In order to develop relevant policies and recommendations tailored according to the needs of immigrant communities, it is essential to understand the interactions between the users within and across sub-communities. Using a novel method (tensor analysis), we reveal the social network structure of an online multilingual discussion forum which hosts an immigrant community in the Netherlands. In addition to the network structure, we automatically discover and categorize monolingual and bilingual sub-communities and track their formation, evolution and dissolution over a long period of time.

Categories and Subject Descriptors

H.2.8 [**DATABASE MANAGEMENT**]: Database applications—*Data mining*; H.4.3 [**INFORMATION SYSTEMS APPLICATIONS**]: Communication Applications—*Bulletin boards*; J.5 [**ARTS AND HUMANITIES**]: Linguistics

Keywords

multilingual, social networks, internet

1. INTRODUCTION

Speaking more than one language is more common than speaking a single language for millions of speakers around the world [4]. Bilingual speakers switch across language boundaries or mix them, depending on their communication partners, the topic of conversation and social factors in context [2]. Similar to face-to-face communication, bilingual speakers also switch across languages in online environments.

Human mobility through immigration is one of the main reasons for contact between speakers of different languages. We perceive multilingualism within immigrant communities as an enrichment to the host community rather than as an obstacle. It is essential for the policy makers to understand the social dynamics, topics of discussion, needs and expectations of the immigrant communities in order to make well-informed decisions and provide them with better recommendations (e.g. education, health, jobs). In this study, we propose a large scale social network analysis of an online multilingual community using a novel analytic method (i.e. tensor analysis).

More specifically, we focus on an online immigrant Turkish community in the Netherlands. Turkish has been in contact with Dutch due to labor immigration since the 1960s and the Turkish community is the largest minority group (2% of the whole population) in the Netherlands [5]. Although the first generation did not learn Dutch beyond the basic level, second and third generations speak both Turkish and Dutch [7] [8] [9]. In addition to Turkish and Dutch, Arabic and English occur occasionally in communication as well.

We focus on the three aspects of social interaction in this study:

- **Discovering Sub-communities**: In addition to the pre-determined sub-forums created by moderators of the online forum, are there also sub-communities that emerge naturally? What are the common factors which bring these sub-communities together?
- **Communication Patterns**: Within each sub-community, what are the emerging social interaction patterns?
- **Temporal Evolution**: Sub-communities are dynamic and their evolution (from formation to dissolution) is triggered by a variety of temporal events. How does the temporal aspect relate to the evolution of sub-communities?

Our contributions are as the following:

- We introduce a novel method based on Tensor Analysis to model and analyze the multilingual communication in an immigrant online community.
- We identify the latent sub-communities which are not directly visible by pre-determined sub-forums created by moderators.
- We analyze the largest (in size) and longest (in duration) multilingual online dataset for immigrant communities so far.
- We leverage automatic language identification in order to minimize human intervention throughout our analysis.

The rest of the paper is structured as follows: Section 2 outlines related work on multilingual communities. Section 4 provides a sketch of our proposed methodology, while Section 3 gives a concise overview of our data set. Section 5 shows our preliminary evaluation on our data set. Finally, we provide future directions and conclude our discussion in in Section 6.

WWW 2015 Companion, May 18–22, 2015, Florence, Italy.
ACM 978-1-4503-3473-0/15/05.
http://dx.doi.org/10.1145/2740908.2743004.

2. RELATED RESEARCH

Multilingual communication has been widely investigated using small scale and mostly spoken data in Sociolinguistics [3] [22] [19]. In immigrant communities, it is very common to mix languages and/or switch across languages depending on the context and the communication parties involved. Each language contact situation is unique and it has been hard to reach a consensus about the terminology, constraints and types of language mixing [28][24][23]. Social factors are predicted to influence the communication between multilingual speakers more than the linguistic factors [26] [11]. Multilingual conversations have also been qualitatively analyzed in online environments [1][27][13] from various aspects. Some studies suggest a link between the topic of discussion and the language choices of multilingual communities [14], [1] and [25].

There is a growing interest within computational areas of research towards automatic language identification for multilingual texts [20] [18] [16] and compilation of mixed language corpora (e.g. [6] for Arabic). More recently, there are also studies focusing on the social aspects of communication between the members of multilingual social networks [15],[10], [12]. These studies mostly make use of Twitter data which differ from real life communication in terms of style and length of conversations.

In contrast to the prior work, we analyze a non-Twitter data set (i.e. online discussion forum) which is closer to informal spoken communication in real life. Our proposed methodology enables us to track the temporal evolution of the online community and the topics that are discussed in sub-communities that are not directly visible.

2.1 Challenges

Throughout this work, we face the following modelling and computational challenges:

- We extract social networks through the comments posted on the online discussion forum. However, social media texts often include unconventional spellings.
- We need automatic language identification for topic modeling and building language profiles of users.
- We seek to extract dense sub-communities of users who post specific tokens to a subset of subforums. Our data are already very sparse. When we introduce time as an additional dimension, the data become even sparser and make such extraction a computational challenge (see Sections 4.3 and 5.3).

3. DATA

The data are extracted from an online forum that mainly hosts Turkish immigrants living in the Netherlands. The forum is online for fifteen years and has pre-determined sub-forums (28) on a variety of topics (e.g. fashion, technology, food). Within the sub-forums, moderators or users pose questions (thread title) for each thread. English and Arabic are used for less than 1% for borrowed words and fixed multi-word units. The data we record spans across 2741 days (from June 2005 until December 2012). The total number of users is over 14 thousand and the total number of posts throughout this period of time is over 4.5 million.

We used an automatic language identifier to label the language of the words in posts. In particular, we employ the technique used in [20] that achieves a word accuracy of 97% ([20] compares different techniques, however we chose the one that achieved the higher word-level accuracy).

4. METHODOLOGY

Our method is based on Tensor Analysis, a very powerful analytical tool that has a plethora of applications [17]. First, we will provide an introduction to tensors and tensor decompositions. Secondly, we describe how we model a multilingual online community within this framework. Finally, we sketch our tensor based algorithms in order to 1) discover latent bilingual sub-communities automatically, 2) infer communication ties between members of each (sub)-community, 3) track the temporal profile of the hidden sub-communities. Tensors are denoted as $\underline{\mathbf{X}}$, matrices as $\mathbf{X}$, and vectors as $\mathbf{x}$. A n-mode tensor is an n-dimensional matrix (e.g. a matrix is a two mode tensor, and a data cube is a three mode tensor). A detailed outline of tensor decompositions are available in [17]. For the current study, we focus on the Canonical (PARAFAC decomposition), which is a simple, easy to interpret and highly used model in this area of research:

$$\underline{\mathbf{X}} \approx \sum_{f=1}^{F} \mathbf{a}_f \circ \mathbf{b}_f \circ \mathbf{c}_f$$

where $\mathbf{a} \circ \mathbf{bc}$ is a rank-one tensor (or a rank-one component of the decomposition) and its (i, j, k)-th element is equal to $\mathbf{a}(i)\mathbf{b}(j)\mathbf{c}(k)$. We represent the decomposition as a set of three matrices $\mathbf{A}, \mathbf{B}, \mathbf{C}$ whose f-th column is $\mathbf{a}_f, \mathbf{b}_f, \mathbf{c}_f$ respectively.

4.1 Automatic Discovery of Latent Sub- Communities

Currently, sub-forums of a variety of topics are pre-determined sub-forums by the moderators and dedicated to the discussion of a variety of topics (e.g. fashion, food) in a top-down manner. However, these sub-forums are often very general and do not instantly reveal the sub-communities that are dynamic and evolve over time. Instead, we follow a bottom-up data driven approach to reveal the latent sub-communities.

Our definition of sub-community is *a set of users, posting consistently a certain set of tokens on a particular set of sub-forums*. A *token* can be anything in a post (e.g. a single term, a smiley/emoticon, a multi-word expression, even a link to external content). Our methodology is generic enough to support a great variety of tokens. Therefore, each sub-community has a notion of a latent discussion topic and a dominant language associated with that topic. The dominant language emerges through prominent tokens used by the particular sub-community.

By accommodating a general set of tokens that a user can post (e.g. emoticons, links etc), we face the challenge of assigning a language to a token that may not be a simple word. If a token refers to a link of external content, the language of that token is inferred via the dominant language of the external content. In the case of emoticons, we encounter two cases: some emoticons transcend language, e.g. `:-)` and some emoticons are specific to a particular language, e.g. `[smiley:liefde]` (where "liefde" means "love" in Dutch) which is a valid emoticon code in our online forum. For the purposes of the present work, we use the language tag that a token has received during the automatic language recognition step. An alternative choice would be to assign a neutral language label to such tokens. We leave analyzing the sensitivity of each particular choice for future work.

Sub-communities as tensor components.

Figure 1 illustrates how we model and analyze the forum data into sub-communities. As a first step, we model the forum as a three mode tensor, where the modes are (user, subforum, token); thus, the (i, j, k) value of tensor $\underline{\mathbf{X}}$ indicates how many times user

i posted token k on subforum j. Note that we start with a static representation of the forum to establish our methodology , and later incorporate time as an additional dimension (see Section 4.3).

Given tensor $\underline{\mathbf{X}}$, we analyze it using the PARAFAC decomposition and obtain a sum of rank-one tensors (i.e. a sum of triplets of vectors). Vectors $\mathbf{a}_f, \mathbf{b}_f, \mathbf{c}_f$ within a triplet are *latent embeddings* that correspond to users, sub-forums, and tokens respectively. Each latent dimension of the embeddings defines a sub-community, and the values of the embeddings declare the membership of each user, sub-forum and token associated with each sub-community. An intuitive way to interpret these embeddings is the following: High values on $\mathbf{a}_f$ refer to users with frequent participation in a certain sub-community f. Accordingly, vectors $\mathbf{b}_f$ and $\mathbf{c}_f$ are interpreted similarly.

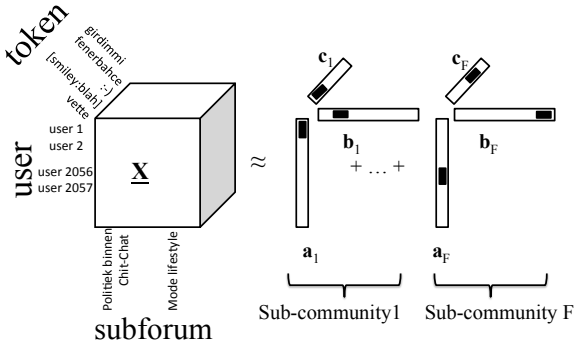


Figure 1: Pictorial example of decomposing the forum into sub-communities by using the CP/PARAFAC decomposition.

Language Profiles of Sub-communities.

Representing language profiles of the sub-communities in an unsupervised way gives us invaluable insights about what holds the community together. We identify the main language of the latent topics discussed within each sub-community automatically based on the tensor decomposition of the online forum. The decomposition produces latent embeddings for the users, sub-forums and tokens.

The latent embeddings of the tokens that describe the latent topic of a particular sub-community are the $\mathbf{c}_f$ vectors, as shown in Fig. 1. For the f-th sub-community, the nonzero values of vector $\mathbf{c}_f$ serve as membership coefficients of tokens to the latent topic of discussion in the sub-community. We use a fully automatic method to tag every post based on the language of the tokens. In this way, we simply count the percentage of Turkish and Dutch tokens that belong to sub-community f and assign a language label to the sub-community. Depending on the particular mixture of the two languages, we label a sub-community as mostly Turkish, mostly Dutch or bilingual (Turkish and Dutch). We do not define a clear cut-off for bilingual versus monolingual sub-communities. If the Turkish and Dutch profiles of a sub-community are within about 10% then the sub-community is most likely to be bilingual. As the percentage of a language decreases, the profile of the sub-community shifts to monolingual.

Language Profiles of Users.

We determine the language profiles by assigning language labels to the sub-communities but also to the individual users based on their level of participation in each sub-community. More specifically, we compute the percentage of Turkish and Dutch she uses as a weighted sum of her participation to all F sub-communities (given by the user to sub-community embedding matrix $\mathbf{A}$) and the Turkish and Dutch labels of those sub-communities, respectively. Determining the individual user profiles are helpful for developing recommendation systems that are customized to the needs and preferences of a particular user or the sub-communities that have members with similar profiles.

One straightforward way to achieve this goal is exhaustively enumerating all tokens posted by a user and measure the percentage of Turkish and Dutch words she uses. However, using the summary provided by the tensor decomposition has the following advantages:

- We can intuitively understand the language label of a user based on her participation in a variety of sub-communities. For instance, if we know that a certain user participates by 80% to an entirely Turkish sub-community and by 20% to an entirely Dutch one, then this user uses 80% Turkish and 20% Dutch by our definition. Looking back, we can argue why a user was profiled this way. Without the notion of a sub-community (which is introduced by our method), such intuitive explanation of a user's language profile is not possible.
- We use automatic language tagging, which is remarkably accurate, albeit not perfect. Automatic summarization of the data helps us to highlight the most prominent activity in the forum and underplays less frequent activity which is more prone to mislabelling.

Language Profiles of Subforums.

The online forum already contains a pre-defined set of subforums (28), roughly divided according to the topic of discussions. However, our sub-communities span across more than one subforum. Therefore, it is worth investigating the language profile of a subforum given the language profiles of the sub-communities that it is part of. We calculate the language profile per subforum the same way we did for users but we now use the latent embeddings of subforums to sub-communities (i.e. matrix $\mathbf{B}$ that stores the latent vectors $\mathbf{b}_f$ of Fig. 1).

4.2 Inferring Communication Patterns & Uncovering the Hidden Social Network

Given the decomposition, we infer high degrees of communication between certain users.

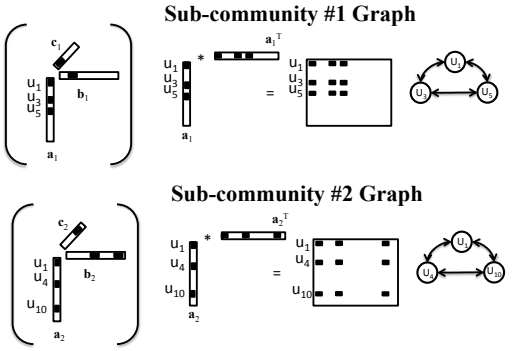


Figure 2: Inferring the hidden communication patterns

Figure 2 illustrates how we uncover the communication patterns between users at the sub-community level by using the *user embed-*

dings, i.e. vectors $\mathbf{a}_f$ as shown on Fig. 1. The non-zero values of vector $\mathbf{a}_f$ indicate the users who participate in the particular sub-community. In order to obtain the adjacency matrix of the graph between those users, we multiply $\mathbf{a}_f\mathbf{a}_f^T$, as shown in Fig. 2 and obtain a user-by-user matrix.

We apply the same procedure to all F sub-communities, deriving graphs that capture the communication patterns within each sub-community. Subsequently, we use the local graphs to synthesize the inter sub-community graph, that captures the interactions across different sub-communities. To do so, it suffices to sum up all local adjacency matrices into a single adjacency matrix. Since the result is not always sparse, we set very small values on the adjacency matrices to zero, deciding the cut-off point automatically. We do 2-means clustering on a vector that contains the non-zero values of the matrix. Then, we take the cluster with the maximum mean and choose the cut-off value to be equal to the smallest value in that cluster, such that any value below that threshold is zeroed out. In this way, we avoid interpreting the noise as connections between users in the same sub-community or across sub-communities.

4.3 Tracking the Sub-Communities over Time

So far, we described how we model the online multilingual community as a three mode tensor, (user, sub-forum, token as the modes). However, sub-communities are not static. On the contrary, we observe various temporal profiles for each sub-community. For instance, some sub-communities are persistent over time, when the topic of discussion pertains to a relatively (temporally) stable topic. Whereas other sub-communities are seasonal (when their underlying topic recurs periodically: the fasting month of Ramadan every year, UEFA's Champions League for football), or even singular time events that last for a certain period of time and then disappear forever.

Our analysis sheds light to this temporal aspect. When we introduce a fourth mode (i.e. date) to our tensor in addition to the existing ones (user, subforum, token), we extend the PARAFAC decomposition, introducing factors $\mathbf{d}_f$ that capture the temporal profile of each one of the f sub-communities. More specifically,:

$$\underline{\mathbf{X}} \approx \sum_{f=1}^{F} \mathbf{a}_f \circ \mathbf{b}_f \circ \mathbf{c}_f \circ \mathbf{d}_f$$

and, as previously, we represent the temporal profiles of all F sub-communities as the columns of matrix $\mathbf{D}$. By augmenting our analysis with a fourth mode, we introduce a new challenge that is not obvious at first sight. Suppose that our data have a certain number of observed (non-zero) entries, which are distributed across the three modes of (user, sub-forum, token). When we introduce the fourth mode, we keep the number of observed entries constant, while increasing the space that these entries can occupy. This results in greater sparsity of data which were already highly sparse and poses challenges for the extraction of coherent components.

5. EXPERIMENTAL RESULTS

From the forum data, we create two different tensors: a three mode (user, subforum, token) tensor of size $14593 \times 28 \times 3053917$ (with density 7×10^{-5}) and a time-evolving (user, subforum, token, day) of size $14593 \times 28 \times 3053917 \times 2741$ (with density 3×10^{-8}). We define the density as the number of non-zero entries in the data, divided by the product of the dimension (i.e. a fully dense tensor will have density equal to 1). As expected, the data we are dealing with are high dimensional and highly sparse. We use ParCube [21], a fast and scalable tensor decomposition algorithm that handles data of this volume efficiently.

5.1 Automatic Discovery of Latent Sub- Communities

First, we show results on the (user, subforum, token) tensor, extracting 15 sub-communities. Table 1 is an indicative subset of the extracted sub-communities (we omit the user-names for clarity and for privacy issues). One example of a bilingual sub-community is #1 of Table 1, where the underlying topic of discussion seems to be revolving around soccer (popular soccer teams such as `galatasaray` and `fenerbahce` appear).

In addition to these sub-communities, Figure 3 contains the language profiles of (a) sub-communities, (b) users, and (c) subforums. We observe a variety of sub-communities (in terms of language profile) spanning across the entire spectrum of almost entirely monolingual to bilingual (with nearly equal participation of both Turkish and Dutch). This is in line with the previous literature which suggests that multilingual users utilize different languages or mix them for different topics of discussion [25].

5.2 Inferring Communication/Social Ties

Figure 4 is a so-called *spy-plot* of the user-to-user interaction adjacency matrix obtained through our analysis of the forum to 15 sub-communities. The rows and columns of the figure correspond to users of the forum, and a blue dot between the i-th row and the j-th column represents an inferred connection between those users.

The upper part of the graph (roughly corresponding to about 2000 users) shows high degrees of activity and dense interconnection between users. The rest of the users (which are the majority) show very faint, close to none, interconnection. In other words, the user communication is highly skewed in this forum. It is common for many users in social media to create an account and rarely use it for various possible reasons (e.g. losing interest, losing log-in credentials, creating fake or secondary accounts for activities like *trolling*). On the other hand, there is an an active subset of users who are in the core of the community and engage in daily conversation with each other. The next step is to identify these highly active users and develop recommendation systems targeting their needs and expectations.

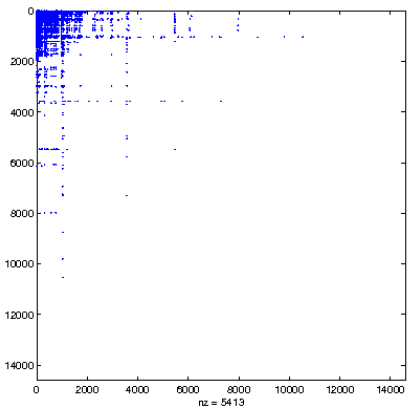


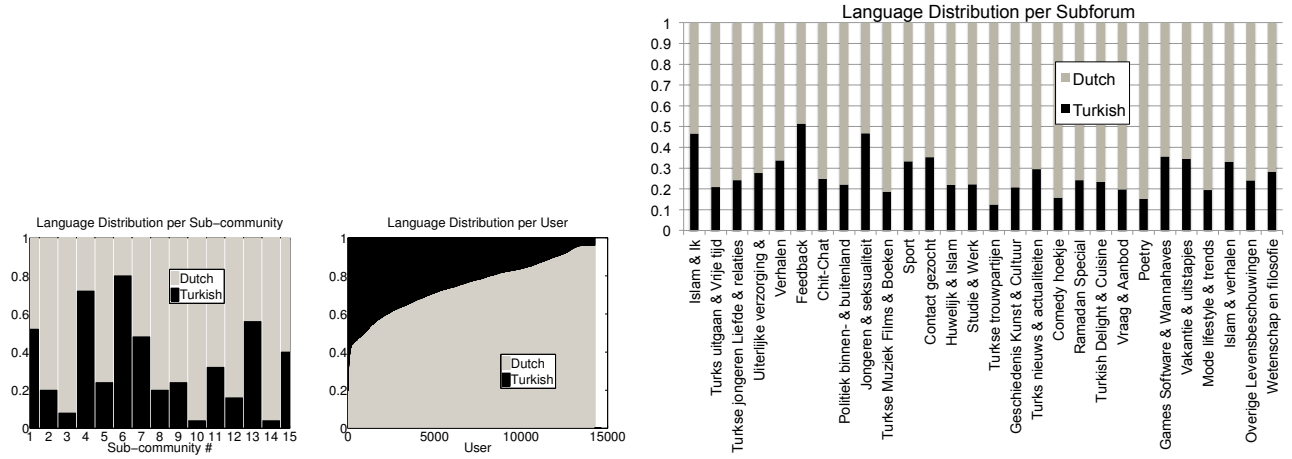
Figure 4: Inferred network between users with strong communication ties

5.3 Tracking the Sub-Communities over Time

To track down the temporal profile of sub-communities, we introduced the "day" mode, which is included in the four mode (user,

Sub-community #	Subforums	Tokens	Language Distribution (TR/Dutch)
1	Jongeren & seksualiteit Politiek binnen- & buitenland Islam & Ik	[smiley:grin], **een**, **je**, :), galatasaray, **als**, **te**, yok, iyi, [smiley:bleh], sonra, **wat**, cok, fenerbahche, mi, gol, adam, bi, degil, son, icin, ilk	0.52 / 0.48
4	Verhalen Feedback Islam & Ik	[smiley:grin], allah, hz, icin, sonra cok, yok, ebu, [smiley:bleh], mi, bi, degil, :) sana, zaman, iyi, kabul, sezon, milli, guzel, hadis	0.72 / 0.28
8	Islam & Ik Mode lifestyle & trends Politiek binnen- &	**je**, **een**, **te**, **als**, ajanslar, **wat**, **mensen**, bakanÃś **politie**, **worden**, **moslims**, **turkije**, **wordt**, buitenland, deze bildirildi, **twee**, **hier**, **veel**, turkmenlerin, **gaat**, **doen**, **land**. kaydetti, **tegen**	0.2 / 0.8
10	Turks uitgaan & Vrije tijd Islam & Ik Chit-Chat	**je**, **een**, **te**, **wat**, [smiley:grin], [smiley:bleh], **echt**, **ga**, **mensen**, **weet**, **hier**, **doen**, **wil**, **goed**, **mn**, **man**, **nee**, **vind**, **gaan**, **gaat**	0.04 / 0.96

Table 1: This table illustrates 4 different sub-communities with different language distributions. Top to bottom: bilingual, mostly Turkish, mostly Dutch, almost exclusively Dutch. For each sub-community we show the participating subforums (in ranked order), and the top tokens.



(a) Per Sub-community (b) Per User (c) Per Subforum

Figure 3: Language profiles as computed using our method.

subforum, token, day) tensor. This mode records the day when a particular user posted a token in a particular subforum. Including time as a mode is very crucial to our analysis but it comes with a price. When we include it, the density of the tensor drops from 7×10^{-5} to 3×10^{-8}. The number of non-zero entries remains constant but the dimensions grow. This increased sparsity makes the extraction of dense sub-communities a computationally harder problem. Although ParCube [21] can handle this, we do not obtain the exact same sub-communities as in the three mode case. This is an unavoidable side effect of the increased sparsity.

In Figure 5, we show the temporal profile of a sub-community on a daily basis. In particular, the sub-community, which our method profiled as bilingual, appears to be active the entire life-time of the forum, exhibiting spikes of activity that appear to have periodic behavior. Via manual inspection of the tokens, we conclude that the discussion topic is religion and the activity of the sub-community peaks around the month of Ramadan every year. This is a single example of temporal behavior that we encountered. Other examples contain single spikes of activity which indicate the formation and ending a sub-community during the lifetime of the forum. We reserve further investigation of such temporal dynamics in detail for future work.

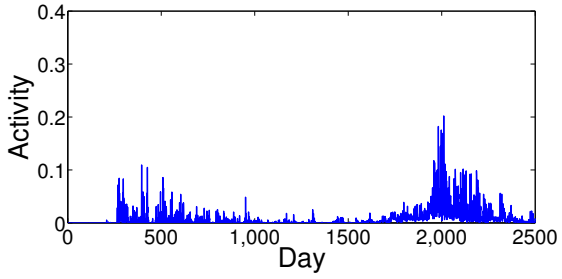


Figure 5: Temporal profile of a bilingual sub-community discussed in Section 5.3

6. CONCLUSIONS

We have presented preliminary work on modelling and analyzing large scale, online multilingual communities in an immigrant setting. We propose a novel and effective methodology that automatically identifies the sub-communities within an online multilingual community, classifies them per language usage, discovers the hidden, underlying social networks between the users who participate in discussions and track the formation, evolution, and dissolution of sub-communities over time. We apply our method to the largest and longest in duration multilingual online forum data set, showing promising and intuitive results of our proposed methodology.

Although preliminary, our results are encouraging for recommendation systems and policy making. Data-driven approaches are especially necessary to assist policy makers in making well-informed decisions for immigrant communities. Future studies are welcome to analyze the communication patterns in other immigrant communities around the world to make comparisons and learn from the practices of each other.

7. ACKNOWLEDGMENTS

The first author was supported by the National Science Foundation Grant No. IIS-1247489. The second author was supported by Digital Humanities Research grant from Tilburg University (Netherlands). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding parties.

8. REFERENCES

[1] J. Androutsopoulos. *The Multilingual Internet*, chapter Language choice and code-switching in German-based diasporic web forums, pages 340–361. Oxford University Press, 2007.

[2] P. Auer. A conversation analytic approach to code-switching and transfer. *Codeswitching: Anthropological and sociolinguistic perspectives*, 48:187–213, 1988.

[3] P. Auer. *Code-switching in conversation: Language, interaction and identity*. Routledge, 2013.

[4] P. Auer and L. Wei. *Handbook of multilingualism and multilingual communication.*, chapter Introduction: Multilingualism as a problem? Monolingualism as a problem, pages 1–14. Berlin: Mouton de Gruyter, 2007.

[5] Centraal Bureau voor de Statistiek. Bevolking, generatie, geslacht, leeftijd en herkomstgroepering. 2013., 2013.

[6] R. Cotterell, A. Renduchintala, N. Saphra, and C. Callison-Burch. An algerian arabic-french code-switched corpus. In *LREC*, 2014.

[7] A. S. Doğruöz and A. Backus. Postverbal elements in immigrant Turkish: Evidence of change? *International Journal of Bilingualism*, 11(2):185–220, 2007.

[8] A. S. Doğruöz and A. Backus. Innovative constructions in Dutch Turkish: An assessment of ongoing contact-induced change. *Bilingualism: Language and Cognition*, 12(01):41–63, 2009.

[9] A. S. Doğruöz and S. T. Gries. Spread of on-going changes in an immigrant language Turkish in the Netherlands. *Review of Cognitive Linguistics*, 10(2), 2012.

[10] I. Eleta and J. Golbeck. Bridging languages in social networks: How multilingual users of twitter connect language communities? *Proceedings of the American Society for Information Science and Technology*, 49(1):1–4, 2012.

[11] P. Gardner-Chloros and M. Edwards. Assumptions behind grammatical approaches to code-switching: When the blueprint is a red herring. *Transactions of the Philological Society*, 102(1):103–129, 2004.

[12] S. A. Hale. Global connectivity and multilinguals in the twitter network. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*, pages 833–842. ACM, 2014.

[13] V. Hinnenkamp. Deutsch, Doyc or Doitsch? Chatters as languagers–The case of a German–Turkish chat room. *International Journal of Multilingualism*, 5(3):253–275, 2008.

[14] J. W. Y. Ho. Code-mixing: Linguistic form and socio-cultural meaning. *The International Journal of Language Society and Culture*, 21, 2007.

[15] S. Kim, I. Weber, L. Wei, and A. Oh. Sociolinguistic analysis of twitter in multilingual societies. In *Proceedings of the 25th ACM Conference on Hypertext and Social Media, HT*, volume 14, pages 243–248, 2014.

[16] B. King and S. P. Abney. Labeling the languages of words in mixed-language documents using weakly supervised methods. In *HLT-NAACL*, pages 1110–1119, 2013.

[17] T. Kolda and B. Bader. Tensor decompositions and applications. *SIAM review*, 51(3), 2009.

[18] M. Lui, J. H. Lau, and T. Baldwin. Automatic detection and language identification of multilingual documents. *Transactions of the Association for Computational Linguistics*, 2:27–40, 2014.

[19] C. Myers-Scotton. *Social motivations for codeswitching: Evidence from Africa*. Oxford University Press, 1995.

[20] D. Nguyen and A. S. Doğruöz. Word level language identification in online multilingual communication. In *Proceedings of EMNLP 2013*, 2013.

[21] E. E. Papalexakis, C. Faloutsos, and N. D. Sidiropoulos. Parcube: Sparse parallelizable tensor decompositions. In *Machine Learning and Knowledge Discovery in Databases*, pages 521–536. Springer, 2012.

[22] S. Poplack. Sometimes i'll start a sentence in spanish y termino en espanol: toward a typology of code-switching1. *Linguistics*, 18(7-8):581–618, 1980.

[23] S. Poplack, D. Sankoff, and C. Miller. The social correlates and linguistic processes of lexical borrowing and assimilation. *Linguistics*, 26(1):47–104, 1988.

[24] S. Romaine. Bilingualism (2nd edn). *Malden, MA: Blackwell Publishers*, 1995.

[25] D. Tang, T. Chou, N. Drucker, A. Robertson, W. C. Smith, and J. T. Hancock. A tale of two languages: strategic self-disclosure via language selection on facebook. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work*, pages 387–390. ACM, 2011.

[26] S. G. Thomason and T. Kaufman. *Language contact*. Edinburgh University Press Edinburgh, 2001.

[27] L. Tsaliki. Globalization and hybridity: the construction of greekness on the internet. *The Media of Diaspora, Routledge, London*, 2003.

[28] L. Wei. *Codeswitching in conversation: Language, interaction and identity*, chapter The 'why' and 'how' questions in the analysis of conversational codeswitching, pages 156–176. Routledge, 1998.