GeoTense: Spotting Patterns in Geo-Social Composite Networks with Tensors

Evangelos E. Papalexakis*, Konstantinos Pelechrinis+ and Christos Faloutsos*
*Department of Computer Science, Carnegie Mellon University
+School of Information Sciences, University of Pittsburgh

***Abstract* - Geo-social networks or Location-based Social Networks (LBSNs for short) provide an abundant source of information for human behavior in real-space. The digital trails that LBSN users leave on these systems, mainly through the action of *check-in* (i.e., voluntary sharing of their whereabouts), span multiple dimensions such as, social (who is friend with whom?), geographical (where do people go?), temporal (when do people transit in the city), as well as contextual (what do people do in the city?) forming a composite network. It thus becomes crucial to be able to analyze the above information in a holistic way, that is, without focusing on a specific dimension only. For that, we develop GeoTense, which models a geo-social composite network as a tensor. The latter enables us to utilize an arsenal of linear algebra tools for analyzing the underlying information and identifying existing patterns. More specifically, in this work we propose the use of tensor decomposition for spotting general patterns and anomalies in the check-in behavior of users, by jointly analyzing and breaking down the behavior of all users of the LBSN into simple, interpretable latent patterns. Our evaluations showcase the potentials of GeoTense in spotting *interesting* patterns in geo-social networks.**

I. INTRODUCTION

During the last years a new class of digital social networks centered around spatial information has emerged. Such Location-based Social Networks (LBSN for short) tie the virtual and physical space through location information. Navigation in the urban space involves now a new dimension, the social. People can instantly get information about their environment and make decisions based on what exists nearby and what their friends or other users of the system believe. Some systems can also offer Groupon-like deals, providing monetary incentives for users and corporations to adopt their usage. Gaming aspects of the LBSNs can also provide a sense of competition between socially connected people, which increases engagement as existing research has shown [1].

The digital trails that people leave in such systems capture in detail the human urban mobility around a city. *Check-ins*, the action of voluntarily declaring your location in LBSNs, further provide the context in which this mobility emerges (e.g., why do people exhibit this mobility pattern). This context is largely absent from existing literature in human urban mobility. While daily navigation through the urban space has been repeatedly shown to be highly regular (e.g., [2,3]) it is interesting to understand under what social and urban context people deviate from the expected patterns. This becomes even more crucial since deviation from the periodic patterns observed in LBSNs might not always be due to special events/conditions. In particular, given that humans respond to incentives and that the motives for adopting LBSN usage are now extended to the real-world [4], people can be tempted to game the underlying system and generate fake information.

However, the objective of our work is not focused on fake information but is more generic. More specifically, **the goal of this study is to develop a data-driven system**, **GeoTense**, **that will be able to automatically identify irregular behavioral patterns**, that is, anomalies[1], as we call them in the rest of the paper. GeoTense is based on tensor decomposition, for spotting (interesting and irregular) patterns in geo-social networks. In brief, the main features that form the contributions of our work as compared with existing literature in LBSN analysis and modeling can be summarized in the following:

- GeoTense considers multiple dimensions (social, spatial and temporal) of the modeled network simultaneously in order to identify latent patterns.

[1] While anomalies has a negative connotation, we would like to emphasize here that the patterns identified are not necessarily "bad". They can represent for example an irregular pattern observed during a special event. However, this is still an anomaly, since it is out of the ordinary behavior.

- GeoTense is not tied to a specific application (e.g., fake check-in detection, neighborhood clustering etc.) but it is generic.

In the rest of the paper we will briefly review related literature (Section II) and we will present our approach (Section III). Finally, we will discuss our evaluation plan and results (Section IV) followed by a brief discussion on the scope of our work (Section V).

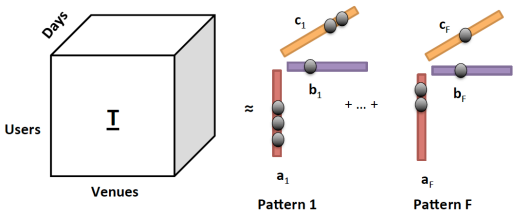

Figure 1. Tensor decomposition in GeoTense.

II. Related Studies

The availability of rich datasets from location-based social networks has lead to a surge in related research. A large volume of the latter focuses on the identification of spatio-temporal patterns crucial for a target application. For example, neighborhood detection and/or characterization is one of the most prevalent applications studied in the area [5-7] (with the list being non-exhaustive). Other studies have focused on the applicability of LBSN data on identifying urban mobility and/or user activity patterns (e.g., [8-10]). Another line of research focuses on the detection of anomalous patterns such as planned and unplanned events (e.g., festivals, traffic accidents etc.). The main high level idea behind the majority of these systems is to detect significant deviations at the volume of the content generated in specific locations [11,12].

Despite the contributions and/or methodological advancements that each one of the above studies exhibit, they are also very focused on a specific, possibly narrow, problem. For example, urban mobility studies are mainly focused on the regular statistical properties of the displacement (i.e., spatial information) of people. The semantics of the locations and/or deviations from these statistics are not considered, while the temporal dimension is many times absent. As another example, event detection schemes are focused on specific type of anomalous patterns. On the contrary, we develop GeoTense without any specific application in mind and hence, it is a generally applicable system. To reiterate, it can be used to identify and further study both regular and anomalous patterns by considering simultaneously the social (users), spatial (locations and their semantics) and temporal (time of check-ins) dimensions.

III. GeoTense in a Nuthshell

We propose to cast the problem of spotting patterns in LBSNs as an instance of composite network analysis using tensors. An n-mode tensor, is a generalization of a matrix (2-mode tensor) in n dimensions, and therefore it forms a natural approach to analyze heterogeneous/composite netwroks. In our case we propose to initially model the spatio-temporal information as a 3-mode (user, venue, time) tensor **T**. Hence, **T**(i,j,k)=1, iff user i was at venue j at time-slot k. Otherwise, **T**(i,j,k)=0.

A typical technique for identifying latent patterns in data represented as a matrix is the Singular Value Decomposition (SVD). A generalization of SVD in n-mode tensors is the *Canonical Polyadic* (CP) or PARAFAC decomposition. In particular, CP/PARAFAC decomposes **T** to a sum of *F* components, such that:

$$\mathbf{T} \approx \sum_{f=1}^{F} a_f \circ b_f \circ c_f \quad (1)$$

where $a_f \circ b_f \circ c_f(i,j,k) = a_f(i) b_b(j) c_f(k)$. In other words, each component (or triplet of vectors) of the decomposition is a rank one tensor. Each vector in the triplet corresponds to one of the three modes of the tensor: **a** corresponds to the users, **b** corresponds to the venues, and **c** corresponds to the days. Figure 1 provides an illustrative example of the tensor decomposition in GeoTense.

One of the issues that we need to consider is how many components we should keep in the sum of Equation (1). Ideally, we would like F <<rank($\mathbf{T}$) for two reasons: 1) we need to project the data to a very low rank subspace, such that similarities across the redundant dimensions will manifest and be expressed in this embedding, 2) identifying the rank of a tensor, unlike the matrix case, is NP-hard. Selecting the appropriate number of components is still an active research topic. For our purposes, we will consider heuristics such as setting a threshold on the number of components, when a certain percentage of the data has been modeled.

Intuition behind the use of tensors: An explanation as to why we expect this proposed approach to offer insightful results is worthwhile. In essence, tensor decompositions attempt to summarize the given (network) data tensor into a reduced rank representation. On the way of accomplishing that, PARAFAC tends to favor dense groups that associate all three entities involved in our data (users, check-ins, and time). These groups need not be immediately visible via inspection of the three mode tensor, since PARAFAC is not affected by permutations of the mode indices. As an immediate outcome of this process, we expect near-bipartite cores (in three modes) of people who check-in at certain places for a certain period of time, to appear as a result of the decomposition, starting from the most dense of them, all the way to the sparsest (if we assume that the rank-one components of the decomposition are sorted by some indicator of density, such as the norm of the three vectors).

Algorithm 1: Spotting anomalies in LBSN using Tensors

Input: User×Venue×Day Tensor $\underline{\mathbf{T}}$, number of group/rank-one components F
Output: F groups $\{\mathbf{A}, \mathbf{B}, \mathbf{C}\}$(as in Fig. 1), and vector l of labels for each group.

Fit a power-law on the degree of the venues in $\underline{\mathbf{T}}$. Let g be the goodness of fit.
Get $\{\mathbf{A}, \mathbf{B}, \mathbf{C}\}$ =PARAFAC($\underline{\mathbf{T}}$,F)
Create label vector l. Default value is 1 (normal component).
for $i = 1 \cdots F$ **do**
  $\underline{\mathbf{G}} = \underline{\mathbf{T}}$
  Let $\mathcal{U}_i$ be the set of users with non-zero weights in the i-th group.
  Set $\underline{\mathbf{G}}(\mathcal{U}_i, :, :) = 0$.
  Fit a power law on the degree of the venues in $\underline{\mathbf{G}}$. Let g' be the goodness of fit.
  if $g' > g$ **then**
    Label the i-th group as an anomaly $\mathrm{l}(i) = -1$.
    $\underline{\mathbf{T}} = \underline{\mathbf{G}}$
  end if
end for

IV. EVALUATIONS

Using a large corpus of check-in data from Foursquare [13] we opt to delve into the details of GeoTense's performance. In particular, we first examine the appropriateness of PARAFAC for analyzing the dataset at hand. Not all datasets are amenable to PARAFAC analysis. Hence, we utilize a very elegant diagnostic tool, namely, CORCONDIA [14], that serves as an indicator that the PARAFAC model describes the data well, or whether there is some problem with the model. The diagnostic provides a number between 0 and 100; the closer to 100 the number is, the better the modeling. If the diagnostic gives a low score, this could be caused either because the chosen rank F is not appropriate, or because the data can not be modelled correctly using PARAFAC, regardless of the rank. In order to better understand whether a variation in the CORCONDIA score is due to bad rank choice or due to data structure, in our experiments we gently increase the rank and observe the behavior. Computing CORCONDIA however, is very challenging. For the purposes of evaluating the behavior of our dataset under PARAFAC model, we carefully implemented CORCONDIA using strict sparse storage and tensor operations from

MATLAB's tensor toolbox. By doing so we were able to run the diagnostic in portions of our dataset. In particular, we chose random samples of 100 users and using their set of check-ins and timestamps we formed tensor $\underline{\mathbf{T}}_s$. Subsequently we took a range of low rank PARAFAC decompositions of $\underline{\mathbf{T}}_s$ and computed the CONCORDIA diagnostic. According to the diagnostic, it appears that until F=5, the data admit a good PARAFAC model, whereas increasing the rank leads to over-factoring. Albeit our analysis is based on a small portion of the users, it shows that if the number of extracted components (i.e., F) is a small percentage of users, the obtained model seems to be of good quality.

We further propose and evaluate a heuristic, unsupervised, algorithm for classifying the tensor components as normal or irregular/anomalous. To reiterate, the latter can refer either to components that are generated from malicious users (e.g., bots) or components that capture irregular, planned events (e.g., organized "flashmobs"). Our heuristic is based on realistic assumptions for the underlying data distribution and in particular on the fact that the distribution of users over the venues is expected to be power law [15,16]. Our algorithm – outlined in Algorithm 1 - examines the improvement of the goodness of fit of a power law to the data once the component under examination is removed. If there is an improvement the component is labeled as anomalous. Our experimental results further showcase the applicability of the proposed algorithm. In particular, by analyzing $\underline{\mathbf{T}}$ we were able to observe a set of distinct patterns, each one belonging to a specific type of anomalous or normal behavior. We chose to extract 1000 components. Among them, only 14 were labeled as anomalous. This fact is encouranging, since intuitively, the anomalies should span only a small portion of the variation in the data. In Figure 2, we demonstrate the progressive improvement of the power-law fit, as we gradually remove each one of the 14 anomalies from the data, while Figures 3 and 4 present exemplar normal and anomalous components respectively.

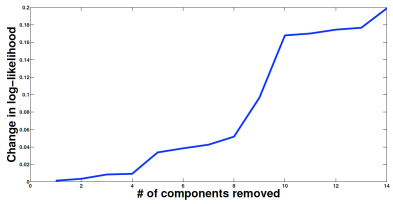

Figure 2. Improvement of the goodness of fit as we gradually remove anomalies.

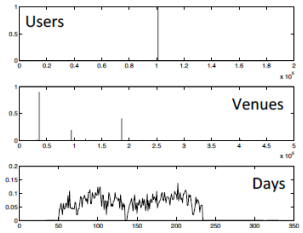

Figure 3. A component labeled as normal from GeoSpot, representing a single user visiting a few venues.

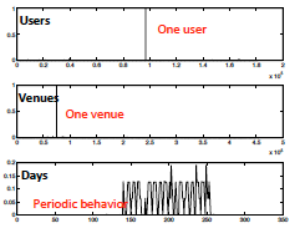

Figure 4. A component labeled as irregular from GeoSpot, resembling the behavior of a bot.

IV. DISCUSSIONS & CONCLUSIONS

One of the things that our evaluations have revealed is the sensitivity of our unsupervised classification algorithm with respect to numerical erorrs at the estimation of the goodness of fit. In particular, there are similar components that barely improve or deteriorate the goodness of the power law fit but due to numerical errors in the estimation they are labeled differently. A *tolerance* factor $\boldsymbol{\varepsilon}$ in the comparison for the goodness of fit could be added to tackle this problem. As another solution, we will also consider other realistic assumptions for the data distribution. Consequently, we will utilize a majority-vote scheme for classifying the components as normal or anomalous.

In summary, in this work we have developed GeoTense, a tensor-based system for spotting patterns in geo-social networks. While our system is generic it can form the basis for more specialized systems (e.g., systems that detect specific attacks possible in LBSNs). However, the most important aspect of GeoTense is that it can lays a framework for analyzing composite networks. Such a framework is largely missing today [17] and we believe that GeoTense can be generalized to multi-dimensional, heterogenous networks.

Acknowledgements: The work has been partially supported by NSF grants IIS-1217559, IIS-1247489, IIS-1408924, CNS-1314632, the ARL grant W911NF-09-2-0053 and the ARO YIP award W911NF-15-1-0599 (67192-NS-YIP).

REFERENCES

[1] A. Anderson, D. Huttenlocher, J. Kleinberg and J. Leskovec, "Steering User Behavior with Badges", in ACM WWW 2013.
[2] C. Song, T. Koren, P. Wang and A.L. Barabasi, "Modelling the Scaling Properties of Human Mobility", in Nature 6, 818-823, 2010.
[3] C. Song, Z. Qu, N. Blumm and A.L. Barabasi, "Limits of Predictability in Human Mobility", in Science, 327(1018), 2010.
[4] J. Lindqvist, J. Cranshaw, J. Wiese, J. Hong and J. Zimmerman, "I'm the Mayor of my House: Examining why People Use Foursquare – a social-driven location sharing platform", in ACM CHI, 2011.
[5] J. Cranshaw, R. Schwartz, J. Hong and N. Sadeh, "The Livehoods project: Utilizing Social Media to Understand the Dynamics of a City", in AAAI ICWSM 2012.
[6] J. Cranshaw and T. Yano, "Seeing a home away from the home: Distilling proto-neighborhoods from incidental data with latent topic modeling", in NIPS workshop on Computational Social Science and the Wisdom of Crowds, 2010.
[7] L. Ferrari, A. Rosia, M. Mamei and F. Zambonelli, "Extracting Urban Patterns from Location-based Social Networks", in ACM LBSN 2011.
[8] A. Noulas, C. Mascolo and E. Frias-Martinez, "Exploiting Foursquare and Cellular Data to Infer User Activity in Urban Environments", in IEEE MDM 2013.
[9] A. Noulas, S. Scellato, R. Lambiotte, M. Pontil and C. Mascolo, "A tale of many cities: universal patterns in human urban mobility", in PLoS ONE 7(5): e37027, 2012.
[10] A. Noulas, S. Scellato, C. Mascolo and M. Pontil, "An empirical study of geographic user activity patterns in Foursquare", in AAAI ICWSM 2011 (poster session).
[11] K. Watanbe, M. Ochi, M. Okabe and R. Onai, "Jasmine: a real-time local-event detection system based on geolocation informantion propagated to microblogs", in ACM CIKM 2011.
[12] R. Lee and K. Sumiya, "Measuring Geographical Irregularities of Crowd Behaviors for Twitter-based Geo-social Event Detection", in ACM LBSN 2012.
[13] Z. Cheng, J. Caverlee, K. Lee and D. Sui, "Exploring millions of footprints in location sharing services", in AAAI ICWSM 2011.
[14] R. Bro and H.A. Kiers, "A new efficient method for determining the number of components in PARAFAC models", in Journal of Chemometrics, 17(5):274-286, 2003.
[15] H. Gao, J. Tang and H. Liu, "Exploring social-historical ties on location-based social networks", in AAAI ICWSM 2012.
[16] G. Lee, S. Rallapalli, W. Dong, Y. Chen, L. Qiu and Y. Zhang, "Mobile Video Delivery in Human Movement", in IEEE SECON 2013.
[17] Y. Sun and J. Han (2013) Mining heterogeneous information networks: a structural analysis approach. ACM SIGKDD Explorations Vol. 14, No. 2, pp. 20-28.