Spotting Misbehaviors in Location-based Social Networks using Tensors

Evangelos Papalexakis
School of Computer Science
Carnegie Mellon University
epapalex@cs.cmu.edu

Konstantinos Pelechrinis
School of Information Sciences
University of Pittsburgh
kpele@pitt.edu

Christos Faloutsos
School of Computer Science
Carnegie Mellon University
christos@cs.cmu.edu

ABSTRACT

The proliferation of mobile devices that are capable of estimating their position, has lead to the emergence of a new class of social networks, namely location-based social networks (LBSNs for short). The main interaction between users in an LBSN is location sharing. While the latter can be realized through continuous tracking of a user's whereabouts from the service provider, the majority of LBSNs allow users to voluntarily share their location, through *check-ins*. LBSNs provide incentives to users to perform check-ins. However, these incentives can also lead to people faking their location, thus, generating false information. In this work, we propose the use of tensor decomposition for spotting anomalies in the check-in behavior of users. To the best of our knowledge, this is the first attempt to model this problem using tensor analysis.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications—*Data Mining*

Keywords

Tensor; Location Based Social Networks; Anomaly Detection

1. INTRODUCTION

Location Based Social Networks tie the virtual and physical space through location information. The latter can enable a number of novel services. Some systems offer Groupon-like deals, providing monetary incentives for users and corporations to adopt their usage. As another example, the LBSN provider can use all the information generated from check-ins to perform recommendations.

Given that humans respond to incentives and that the motives for adopting LBSN usage are now extended to the real-world [5], people are tempted to game the underlying system and generate false information. In fact, as very illustratively posited by LBSNs service providers "unfortunately, most of them [check-ins] are fake" [1]. Therefore, this socio-spatio-temporal information will be devalued as long as it is not trustworthy.

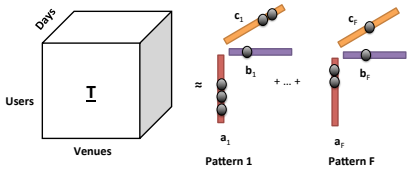


Figure 1: Given a (user, venue, day) tensor $\underline{\mathbf{T}}$, we obtain a decomposition into F rank-one components, each corresponding to a different latent pattern in the data.

In this work, we propose a novel approach, based on tensor decomposition, for spotting anomalies[2] in LBSNs. Our system is generic, in the sense that it does not target any specific misbehavior. We model the problem of detecting *interesting* and potentially misbehaving patterns in LBSNs by formulating it as a tensor, and consequently, analyzing and summarizing the data leveraging tensor decompositions. Each identified component is essentially a *bag of check-ins*, possibly from multiple users, over multiple time-periods.

Related Studies There has been some body of work pertaining to location faking in LBSNs [4], *sybil attack*[3] detection [8]. There has been work in the literature that employs tensors for anomaly detection such as [6, 7]. However, to the best of our knowledge, this is the first attempt to model and tackle the problem at hand using tensor decompositions.

2. OUR APPROACH

We propose to cast the problem of spotting anomalies and misbehaviors in LBSNs as an instance of tensor analysis. An n-mode tensor, is a generalization of a matrix (2-mode tensor) in n dimensions. In our case we model the spatio-temporal information as a 3-mode (user, venue, time) tensor $\underline{\mathbf{T}}$. Hence, $\mathbf{T}(i,j,k) = 1$, iff user i was at venue j at time k. Otherwise, $\underline{\mathbf{T}}(i,j,k) = 0$. A means of analyzing a tensor is the *Canonical Polyadic* (CP) or PARAFAC decomposition [3]. In particular, CP/PARAFAC decomposes $\underline{\mathbf{T}}$ to a sum of F components, such that $\underline{\mathbf{T}} \approx \sum_{f=1}^{F} \mathbf{a}_f \circ \mathbf{b}_f \circ \mathbf{c}_f$, where $\mathbf{a}_f \circ \mathbf{b}_f \circ \mathbf{c}_f(i,j,k) = \mathbf{a}_f(i)\mathbf{b}_f(j)\mathbf{c}_f(k)$ In other words,

WWW'14 Companion, April 7–11, 2014, Seoul, Korea.
ACM 978-1-4503-2745-9/14/04.
http://dx.doi.org/10.1145/2567948.2576950.

[1] http://techcrunch.com/2010/08/03/shopkick-best-buy
[2] In the rest of the paper we will use the term anomaly and misbehavior interchangeably.
[3] A sybil attack is an attack where fake accounts generate fake check-ins

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 (see also Fig. 1) Each of these F components can be considered as a cluster, and the corresponding vector elements as soft clustering indicators. For the purposes of this work, we use the, highly optimized, Tensor Toolbox for Matlab [1].

Intuition behind the use of tensors: Tensor decompositions attempt to summarize the given data tensor into a reduced rank representation., favoring dense groups on the way of accomplishing it. 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.

Data description: In our experiments we use a dataset obtained from Foursquare [2]. The original dataset, includes geo-tagged user generated content from a variety of social media that was pushed to Twitter's public feed between September 2010 and January 2011. Each tweet includes location information in the following format: `<userID, tweetID, text, location, time, venueID>`.

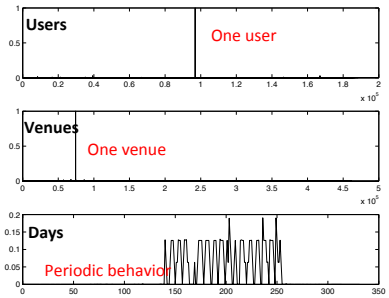
We remove check-ins in locations - i.e., (lat, lon) pairs that can possibly correspond to more than one venues - that have less than 10 check-ins in total and we eventually get our final dataset of 6,699,516 check-ins, in 461,690 venues from 186,083 users. In order to form $\underline{\mathbf{T}}$, we discretize time in bins of one day, and hence, the entry $\underline{\mathbf{T}}(i, j, k)$ of the tensor is the number of check-ins user i made at venue j on day k.

Our results: In Figure 2 we present two exemplary anomalies. The most impressive anomalous pattern discovered is the one of subfigure 2(a), where the temporal evolution of the group of user and venue is almost perfectly periodic, resembling the behavior of a bot.

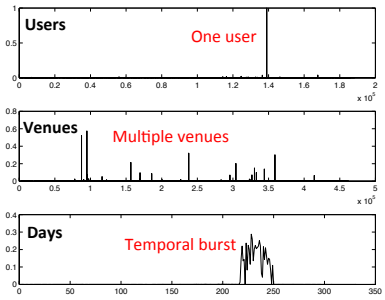
Key to our approach is the temporal dimension. In many cases, the temporal behavior of a group of users and venues is able to differentiate a sybil attack from, say, a group of friends, checking in at the same restaurant at lunch time every day; the sybil attack should exhibit some anomaly in time (either a burst or a spike), whereas the normal behavior, should, usually, be quasi-periodic. By modeling our data as a tensor, and formulating the problem in this fashion, we include time in our analysis, a fact that enables us to produce more intuitive and interpretable results. For instance, the pattern at Figure 2(b) represents a user that has checked-in to a (small) number of venues many times. This is a possible indicator of a user that wants to gain virtual (e.g., points, badges, "mayorships" in Foursquare etc.) and/or real-life rewards (e.g., unlocking special offers that require multiple check-ins) .

3. CONCLUSIONS

In this paper we present a preliminary approach for spotting anomalies in check-in patterns of users in LBSNs using tensors. To the best of our knowledge, this is the first attempt to model this problem using tensor decompositions. Our approach is generic in the sense that it does not require special infrastructure and coordination with the LBSN provider. Future directions include exploiting data sparsity to improve our methods, as well as evaluating our approach on different LBSNs.



(a)



(b)

Figure 2: Anomalous components, as produced by our approach. The most surprising is subfigure (a), where the temporal evolution of the group is perfectly periodic for a period of time, resembling the behavior of a bot.

Acknowledgements

Research was funded by grants NSF IIS-1247489 and ARL under Cooperative Agreement Number W911NF-09-2-0053. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the funding parties.

4. REFERENCES

[1] B.W. Bader and T.G. Kolda. Matlab tensor toolbox version 2.2. *Albuquerque, NM, USA: Sandia National Laboratories*, 2007.

[2] Z. Cheng, J. Caverlee, K. Lee, and D.Z. Sui. Exploring millions of footprints in location sharing services. In *AAAI ICWSM*, 2011.

[3] R.A. Harshman. Foundations of the parafac procedure: Models and conditions for an" explanatory" multimodal factor analysis. 1970.

[4] W. He, X. Liu, and M. Ren. Location cheating: A security challenge to location-based social network services. In *IEEE ICDCS*, 2011.

[5] 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 application. In *ACM CHI*, 2011.

[6] K. Maruhashi, F. Guo, and C. Faloutsos. Multiaspectforensics: Pattern mining on large-scale heterogeneous networks with tensor analysis. In *Proceedings of the Third International Conference on Advances in Social Network Analysis and Mining*, 2011.

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

[8] Bimal Viswanath, Ansley Post, Krishna P Gummadi, and Alan Mislove. An analysis of social network-based sybil defenses. In *ACM SIGCOMM Computer Communication Review*, volume 40, pages 363–374. ACM, 2010.