Conference	Sim_1	Sim_2	Sim_3	Sim_4
ANTS	0.55	0.16	0.14	0.08
TLCA	0.60	0.53	0.14	0.08
INFOS	0.83	0.75	0.10	0.22
Sigite Conf.	0.66	0.33	0.11	0.27
Gil Jahrestagung	0.77	0.81	0.08	0.31

Table 3: Analysis of Top Outlier Conferences (*1Sμ*)

***IMDB*: ** Top two outlier actors returned by *OneStageμ* are discussed below.

1. Kelly Carlson (I): In X_1, she did many Sport, Thriller and Action movies, while in X_2 she switched to Drama, Music, Reality-TV. Most of her top ten closest collaborators in X_1 still do Documentary, Thriller, Action in X_2. Few of her co-stars collaborate with her in the second time snapshot. Also, her X_1 neighbors used to do Sport, Comedy, Documentary, Action while her X_2 neighbors do Drama, Documentary, Thriller, Music. Thus, clearly she changed her community from Sports, Thriller, Action genres to Drama, Music genres.

2. Josh Brolin: In X_1, he did a lot of Thriller, Drama, Crime and Mystery movies. In X_2, he acted in a lot of Documentary, Comedy, History, Music movies. Not only did he change his genres completely, but also not many other actors show such a change in the type of their movies. This is why, in X_2, his genres are quite different from that of his X_1 nearest neighbors. Hence, clearly he is an outlier.

***DBLP* (Authors Network):** We will discuss about the top two authors which are detected as evolutionary community outliers.

1. Georgios B. Giannakis. In X_1, his publications were mainly in CISS, ICC, GLOBECOM, INFOCOM. In X_2, he published in completely different conferences: ICASSP, ICRA. We looked at the conferences at which his X_1 community members published in X_2. This set of conferences (GLOBECOM, ICC, CISS, INFOCOM) is completely different from the set of conferences at which he published in X_2, but much similar to his own published venues in X_1.

2. Vassilios Peristeras. In X_1, his publications were mainly in HICSS, ICEGOV, IEEE SCC, EDOCW, CSREA, etc. In X_2, he published in completely different conferences: WSKS, ICSC, OTM, ICDIM, SAC. We looked at the conferences at which his X_1 community members published in X_2. This set of conferences (HICSS, ISI, ICEGOV, etc.) is completely different from the set of conferences at which he published in X_2, but much similar to his own conferences in X_1. This clearly justifies him to be a community outlier.

***DBLP* (Conf Network):** Table 3 shows the top five conferences returned as outliers by *OneStageμ*. We performed some analysis and hence list four measures in Table 3: similarity between top 20 words (we removed most frequent 100 words from dataset) for the conference across the snapshots, similarity in top 20 words between the conference at X_2 and its ten closest X_1 community members, similarity in neighbor conferences across the two snapshots, similarity in the words shared by the neighbors in the first snapshot and neighbors in the second snapshot.

As the table shows, each of the conferences have very low similarity for at least one of the measures, justifying their detection as *ECOutliers*. For comparison with the baseline, we present the top five conferences returned as outliers returned by *TwoStage* in Table 4. As one can clearly see, the similarity values in Table 4 are much higher compared to Table 3. Thus, the outliers returned by *TwoStage* are not as good as the outliers returned by the proposed algorithm.

Conference	Sim_1	Sim_2	Sim_3	Sim_4
ESEC SIGSOFT FSE	0.26	0.71	0.72	0.70
ISORC	0.64	0.66	0.75	0.74
Communications in Computing	0.16	0.57	0.58	0.70
ICDE Workshops	0.27	0.65	0.74	0.78
HICSS	0.78	0.60	0.83	0.79

Table 4: Analysis of Top Outlier Conferences (*2S*)

***Four Area* (Authors Network):** In this dataset, we observe a trend of people moving from ML community to DM community. Also, many authors often publish in both DB and IR. For this dataset, we will discuss two outliers returned by *OneStageμ*, who behave quite different from these trends.

1. Jérôme Lang. For this author, most of his community members moved from logical reasoning and related areas to other areas like DM and IR but he stays in AI field. He published 5 and 11 papers in the two snapshots respectively. Words in the titles of his papers are mainly "logic, planning, representation, action, uncertainty, propositional". We looked at the words which his X_1 community members (other authors having similar word distributions) use in X_2. The top words were "data, retrieval, xml, web, learning, mining". This clearly shows that his community members moved from logical reasoning to other areas while Jérôme decided to stay in the area, opposed to the trend. While he continued to publish in pure ML and AI conferences, his community members publish in a lot of IR and DM conferences in X_2.

2. Georg Gottlob. Generally the observed trend is that ML authors move to other related areas like DM and IR. Sometimes, some DM authors publish in ML conferences. But Georg has been a DB author in X_1, who started publishing in ML community in X_2. In X_1, he published frequently in PODS, VLDB, ICDE. In X_2, apart from PODS, he published heavily in IJCAI and AAAI. His set of collaborators also changed by a large extent across the two snapshots. A lot of his X_2 collaborators publish in ML conferences.

5.4 Running Time and Convergence

The experiments were run on a Linux machine with 4 Intel Xeon CPUs with 2.67GHz each. The code was implemented in Java. Fig. 4 shows the execution time for *OneStageμ* on different synthetic datasets in ms. Note that the algorithm is linear in the number of objects. These times are averaged across 100 runs of the algorithm. On an average *OneStageμ* needed ∼13 iterations per pass to converge on both real and synthetic datasets. Fig. 5 shows the change in the objective function value with iterations for the *SynMix* dataset for different number of objects, using a log-linear plot. The figure shows that *OneStageμ* converges fairly quickly.

6. CONCLUSIONS

We introduced the notion of evolutionary outliers with respect to latent evolving communities, i.e., *ECOutliers*. Such outliers represent the objects which disobey the common evolutionary trend among the majority of the objects in a community. The challenge is that both community evolution patterns and outliers are unknown. Outliers should be derived based on community matching across different snapshots, but need to be ignored when conducting community matching. We proposed an optimization framework which integrates community matching and outlier detection. The objective function is to minimize community matching error, in which the contributions from outlier objects are weighed lower. An iterative algorithm *OneStageμ* is developed to solve the optimization problem, which improves community matching and *ECOutlier* detection gradually. Experiments on a series of synthetic data show the proposed algorithm's capability of detecting outliers under various types of community evolution. Case studies on *DBLP*, *IMDB* and *Four Area* datasets reveal some interesting and meaningful evolutionary outliers. Although the proposed algorithm focuses on two snapshots, it can detect both short-term and long-term trends and outliers, as snapshots can consist of short or long intervals. Moreover, it can be extended to handle multiple snapshots.

7. ACKNOWLEDGEMENTS

We thank Vinod Vydiswaran and anonymous KDD reviewers for discussions and insightful comments. Research was sponsored in part by the Cyber Security project (W911NF-11-2-0086) of U.S. Army Research Laboratory, NSF IIS-0905215, and U.S. Air Force Office of Scientific Research MURI award FA9550-08-1-0265. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

8. REFERENCES

[1] T. Abeel, Y. Van de Peer, and Y. Saeys. Java-ML: A Machine Learning Library. *Journal of Machine Learning Research*, 10:931–934, Jun 2009.

[2] C. C. Aggarwal and P. S. Yu. Outlier Detection for High Dimensional Data. *SIGMOD Records*, 30:37–46, May 2001.

[3] C. C. Aggarwal and P. S. Yu. Outlier Detection with Uncertain Data. In *Proc. of the SIAM Intl. Conf. on Data Mining (SDM)*, 483–493, 2008.

[4] C. C. Aggarwal, Y. Zhao, and P. S. Yu. Outlier Detection in Graph Streams. In *Proc. of the 27th Intl. Conf. on Data Engineering (ICDE)*, 399–409. 2011.

[5] J. Alon, S. Sclaroff, G. Kollios, and V. Pavlovic. Discovering Clusters in Motion Time-Series Data. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Volume 1, 375–381. IEEE Computer Society, 2003.

[6] D. P. Bertsekas. *Non-Linear Programming (2nd Edition)*. Athena Scientific, 1999.

[7] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander. LOF: Identifying Density-Based Local Outliers. In *Proc. of the 2000 ACM SIGMOD Intl. Conf. on Management of Data (SIGMOD)*, 93–104. ACM, 2000.

[8] V. Chandola, A. Banerjee, and V. Kumar. Anomaly Detection: A Survey. *ACM Surveys*, 41(3), 2009.

[9] W. W. Cohen and J. Richman. Learning to Match and Cluster Large High-Dimensional Data Sets for Data Integration. In *Proc. of the 8th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining (SIGKDD)*, 475–480. ACM, 2002.

[10] E. Dimitriadou, A. Weingessel, and K. Hornik. Voting-Merging: An Ensemble Method for Clustering. In *Proc. of the Intl. Conf. on Artificial Neural Networks (ICANN)*, 217–224. Springer, 2001.

[11] S. Dudoit and J. Fridlyand. Bagging to Improve the Accuracy of a Clustering Procedure. *Bioinformatics*, 19(9):1090–1099, 2003.

[12] E. Eskin. Anomaly Detection over Noisy Data using Learned Probability Distributions. In *Proc. of the 17th Intl. Conf. on Machine Learning (ICML)*, 255–262. Morgan Kaufmann Publishers Inc., 2000.

[13] A. J. Fox. Outliers in Time Series. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(3):350–363, 1972.

[14] J. Gao, F. Liang, W. Fan, Y. Sun, and J. Han. Graph-based Consensus Maximization among Multiple Supervised and Unsupervised Models. In *Proc. of the 23rd Annual Conf. on Neural Information Processing Systems (NIPS)*, 585–593. Curran Associates, Inc., 2009.

[15] J. Gao, F. Liang, W. Fan, C. Wang, Y. Sun, and J. Han. On Community Outliers and their Efficient Detection in Information Networks. In *Proc. of the 16th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining (SIGKDD)*, 813–822, 2010.

[16] Y. Ge, H. Xiong, Z. Zhou, H. Ozdemir, J. Yu, and K. C. Lee. Top-Eye: Top-K Evolving Trajectory Outlier Detection. In *Proc. of the 19th ACM Conf. on Information and Knowledge Management (CIKM)*, 1733–1736, 2010.

[17] A. Ghoting, M. E. Otey, and S. Parthasarathy. LOADED: Link-Based Outlier and Anomaly Detection in Evolving Data Sets. In *Proc. of the 4th IEEE Intl. Conf. on Data Mining (ICDM)*, 387–390, 2004.

[18] V. J. Hodge and J. Austin. A Survey of Outlier Detection Methodologies. *AI Review*, 22(2):85–126, 2004.

[19] W. Hu, Y. Liao, and V. R. Vemuri. Robust Anomaly Detection Using Support Vector Machines. In *Proc. of the Intl. Conf. on Machine Learning (ICML)*, 282–289. Morgan Kaufmann Publishers Inc, 2003.

[20] M. Jakobsson and N. A. Rosenberg. CLUMPP: A Cluster Matching and Permutation Program for Dealing with Label Switching and Multimodality in Analysis of Population Structure. *Bioinformatics*, 23:1801–1806, Jul 2007.

[21] E. M. Knorr and R. T. Ng. Algorithms for Mining Distance-Based Outliers in Large Datasets. In *Proc. of the 24th Intl. Conf. on Very Large Data Bases (VLDB)*, 392–403. Morgan Kaufmann, 1998.

[22] E. M. Knorr, R. T. Ng, and V. Tucakov. Distance-Based Outliers: Algorithms and Applications. *The VLDB Journal*, 8:237–253, Feb 2000.

[23] D. Kottke and Y. Sun. Motion Estimation Via Cluster Matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16:1128–1132, 1994.

[24] H.-P. Kriegel, P. Kröger, E. Schubert, and A. Zimek. LoOP: Local Outlier Probabilities. In *Proc. of the 18th ACM Conf. on Information and Knowledge Management (CIKM)*, 1649–1652. ACM, 2009.

[25] H.-P. Kriegel, P. Kröger, E. Schubert, and A. Zimek. Interpreting and Unifying Outlier Scores. In *Proc. of the 11th SIAM Intl. Conf. on Data Mining (SDM)*, 13–24. SIAM / Omnipress, 2011.

[26] J.-G. Lee, J. Han, and X. Li. Trajectory Outlier Detection: A Partition-and-Detect Framework. In *Proc. of the 24th Intl. Conf. on Data Engineering (ICDE)*, 140–149. IEEE Computer Society, 2008.

[27] B. Long, Z. M. Zhang, and P. S. Yu. Combining Multiple Clusterings by Soft Correspondence. In *Proc. of the 5th IEEE Intl. Conf. on Data Mining (ICDM)*, 282–289. IEEE Computer Society, 2005.

[28] M. J. Miller, A. D. Olson, and S. S. Thorgeirsson. Computer Analysis of Two-Dimensional Gels: Automatic Matching. *ElectroPhoresis*, 5(5):297–303, 1984.

[29] D. Pokrajac, A. Lazarevic, and L. J. Latecki. Incremental Local Outlier Detection for Data Streams. In *IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*, 504–515. IEEE, Apr 2007.

[30] S. Ramaswamy, R. Rastogi, and K. Shim. Efficient Algorithms for Mining Outliers from Large Data Sets. *SIGMOD Records*, 29:427–438, May 2000.

[31] Y. Sun, J. Han, J. Gao, and Y. Yu. iTopicModel: Information Network-Integrated Topic Modeling. In *Proc. of the 9th IEEE Intl. Conf. on Data Mining (ICDM)*, 493–502. IEEE Computer Society, 2009.