Neural Collaborative Filtering to Predict Human Contact with Large-Scale GPS data

Jorge F. Barreras*
Department of Mathematics
University of Pennyslvania
Philadelphia, PA 19104
fbarrer@sas.upenn.edu

Bethany Hsiao
Department of Computer
and Information Science
University of Pennyslvania
Philadelphia, PA 19104
bhsiao@seas.upenn.edu

Hamed Hassani
Department of Electrical &
Systems Engineering
University of Pennyslvania
Philadelphia, PA 19104
hassani@seas.upenn.edu

Duncan J Watts
Department of Electrical &
Systems Engineering
University of Pennyslvania
Philadelphia, PA 19104
djwatts@seas.upenn.edu

Abstract

Understanding and measuring the effect of human mobility on the spread of epidemics is key to addressing these threats. GPS human mobility data represents an enormous advancement in the field of epidemics as it could reveal population-level contact patterns that can replace homogeneous mixing assumptions or the unjustified use of synthetic random networks in epidemic models. However, a standing challenge in the estimation of contacts from GPS signals is addressing the high sparsity in this type of data. Alas, most users are observed only for a small fraction of the time. In this paper, we address this issues by proposing a novel methodology that can fill in the gaps in the data. Our framework is based on link prediction using deep learning to predict missing links in a temporal bipartite graph connecting users and locations. We demonstrate and validate our methodology on privacy-enhanced location data from thousands of mobile devices in the city of Philadelphia during 2020.

1 Introduction

In the past two years, large-scale and highly granular human mobility data [1] have revolutionized networked models of epidemics, enabling them to capture dynamic patterns of human interaction and has been influential in research and policy design alike [CITATIONS]. Yet, mathematical modeling of epidemics integrating this class of data still faces important challenges that limit their usefulness in practice. One of these challenges is the construction of epidemic networks that capture the contact patterns of individuals at the population level.

Research before the COVID-19 pandemic mostly inferred human contact patterns from datasets with coarser resolution (e.g. airline flight seat data or survey commuting data) by making the assumption that individuals come in contact randomly and homogeneously in large subpopulations (e.g. countries or areas serviced by airports) and move randomly between them. In contrast, epidemic models in the literature after the COVID-19 outbreak used GPS human mobility data that is granular enough to

*Corresponding Author.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

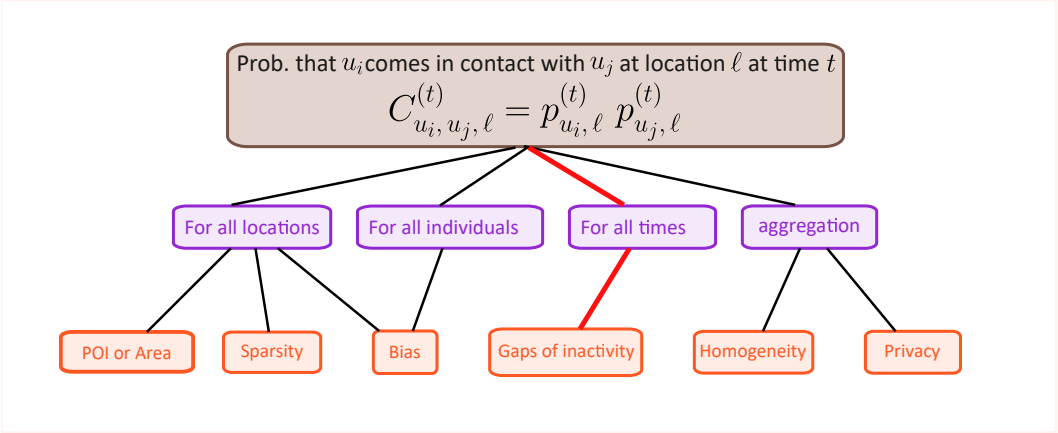

Figure 1: Standing challenges associated with the construction of contact networks from GPS data. In this work we address the issue of gaps of inactivity in user data to construct a bipartite network connecting users and locations for all hours of the week (highlighted in red).

allow estimates of contact at the hourly level with high spatial accuracy—dispensing with assumptions of homogeneity. While human contact can be inferred purely from spatial and temporal proximity [2]; a more common approach is to assume that contact between two individuals only occurs when both are present in the same place of interest (POI), in a certain bucketed time period (e.g. hour or day) [3, 4, 5, 6, 7]. For the latter approach, we denote the probability that agents u_i and u_j come in contact with one another in POI p during time bucket t as $C^{(t)}_{u_i,u_j,p}$. Assuming independence, we can compute this probability as the product of the marginal probabilities that u_i and u_j visit p during t, respectively. Namely, our strategy is to estimate the right hand side of $C^{(t)}_{u_i,u_j,p} = P^{(t)}_{u_i,p} \cdot P^{(t)}_{u_j,p}$.

Estimating such probabilities of contact first requires addressing some important limitations in this type of data—an issue that has received little attention in the recent literature. An initial concern is the definition of the locations in which contacts are assumed to occur; using a database of physical venues or POIs (as in [5, 3, 7]), provides a very accurate estimation of human contact but it will exclude all the contacts occurring in other places not in the data—biasing the estimates if the sample of POIs is small and not representative. Alternatively, using administrative areas can achieve full spatial coverage of a region (as in [6]), but will estimate contact poorly since the areas are too large. Another issue is related to the sample of individuals, which can be small, geographically biased, and can change in composition over time as data from different apps is included or excluded. Thus, normalization is needed when estimating patterns and numbers of visits for the whole population; simple approaches like weighting the number of visitors in a given hour [**?**, 3] will result in very sparse estimates, since the original visitation data can be extremely sparse. A third aspect of dealing with this type of data to construct networks is the level of aggregation: agent-level networks allow for re-identification of individuals [CITATION] and, in addition, might be too large and make epidemic models intractable. Privacy is preserved more when individuals are aggregated into communities (e.g. at the CBG level) then the estimated contact patterns will not capture sub-community heterogeneities. A final data issue that has been mostly overlooked is the sparsity of the data at the individual level; namely, there are large gaps of inactivity in user data likely owing to the passive nature of the data collection. Notwithstanding the possibility that there are measurement biases, accurately estimating contact patterns for all individuals requires a methodology to estimate such missing activity. The methodological requirements just described are summarized in Figure 1. In this paper, we tackle the latter issue by proposing a methodology based on neural collaborative filtering to accurately predict the location of users that are missing in the data.

Our approach bears resemblance to neural collaborative filtering [8, 9], in which a user's preference for an item (in our case a location) is estimated based on the preferences of similar users. This is achieved by means of a fine-tuned neural network that embeds users and items into low-dimensional spaces in which similar entities are embeded close to each other; this neural network then uses such latent representations to predict the preference of users for items. This collaborative filtering problem

can be interpreted as a link prediction problem in a relational graph, in which a user u is connected to a location p with the relation t, if u visits p at time t.

Our approach makes use of a novel approach to generate informative and factual negative samples. We demonstrate our method on real anonymized GPS mobility data from hundreds of thousands of smartphone devices in the city of Philadelphia during 2020.

Our contributions can be summarized as:

- A new methodology applying deep-learning-based link prediction to large-scale GPS data to generate epidemic networks.
- A negative sampling methodology that makes training more efficient by proposing factual and informative negative samples.

2 Methodology

2.1 Location Data

We use GPS human-mobility data from a leading location intelligence company[2] consisting of location data aggregated from millions of devices in the US and sourced using the location services of thousands of mobile apps in which users have consented to location-sharing. The data consists of time-stamped GPS coordinates tied to unique anonymized device identifiers. We have access to a derivative dataset in which pings corresponding to user stops (or *stays*) are clustered together using a variation of the DBScan clustering algorithm [10], which are represented as a centroid and duration. We also identify a individual's weekly home location as the location[3] where they spend most night hours in the preceding 6 weeks, conditional on visiting that location at least three different weeks. Predicting users' visits to locations requires attributing a given venue or location to each stay; while visitation data tied to specific business venues is offered by many providers and is common in location prediction tasks [12], we prefer to join stays at the coarser level of Census block groups (CBGs) to not exclude stays from our analysis, nor bias our results if the sample of POIs is biased. The CBG associated with a user's home stays is further replaced with the token location "HOME". This censoring protects user privacy and reflects the modeling assumption that we are only interested in contacts outside of households (individuals in the same CBG in the same hour are not at risk of infection if they are home).

Our models are trained and tested using GPS mobility data from the city of Philadelphia; i.e. GPS pings that fall within the Census designated urban limits. As an initial step, we cluster pings using a time-augmented implementation of DBSCAN [13] that results in *stays*; clusters of pings attributed with a centroid, a start time and a duration. We further prepare the data by "exploding" each stay into the hours the stay straddles. This results in triples of the form (u, t, ℓ) representing that user u was present at location ℓ during (part of the) hour t. These triples represent a subset of observed edges in a *temporal bipartite graph*. The training is performed on the week of 04/01/2020 - 04/07/2020 and validation is done in the following week corresponding to 04/08/2020 - 04/14/2020. There are 50,403 users in the sample for which there is an identified home location in the training period. However, at the user level there is significant sparsity in the signals of a given user. Figure 2 shows the distribution of a statistic measuring coverage at the individual level; namely, the fraction of the hours of the training week with at least one stay. As it can be observed from this skewed bimodal distribution, over 50% of the users are observed less than 25% of the hours—implying that many visits and spreading events are not observed.

2.2 Location Prediction with Neural Collaborative Filtering

As it was mentioned in the introduction, our main goal is to predict the visits of a given user for every hour of the week using sparse data. Because we only observe a given user for a fraction of the time, we rely on a NCF approach to predict the likely locations visited in the missing hours; effectively learning mobility patterns of a user from the data of similar users. NCF has been proven to be more

[2]Kept anonymous while pending pre-publishing review by the provider.

[3]These locations correspond to geographic tiles known as geohash-7 [11], to which stays are uplifted.

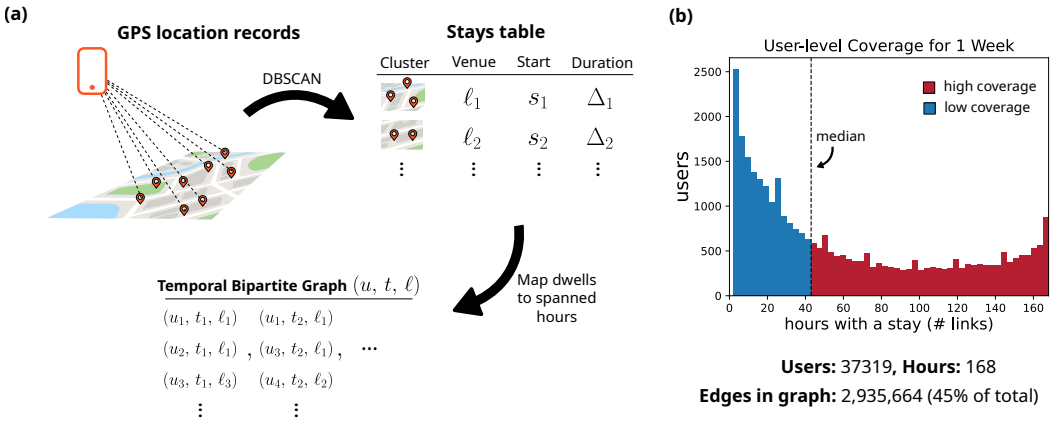

Figure 2: **(a)** Methodology to process location data and construct a temporal bipartite graph. **(b)** Distribution of user-level coverage, defined as the number of hours of the week of 01-04-2020 that have a stay from the user (an observed link in the bipartite graph); the distribution is bimodal and we use the median to divide between high coverage and low coverage users.

versatile than matrix factorization for recommender systems [14, 9]; the latter being constrained by the linearity of the output and the inability to incorporate side information.

For our experiments, we split data from our first week into training and validation sets using an 80-20 cross-validation setup, and we use the second week as held-out test data. For training and validation we construct records of the form (u, i, t) where u is a user, i is a location the user visited, and t is an hour of the week. To this end, we simplify the problem by constructing several records from a single stay which might have an arbitrary duration, into a single record for each hour that the stay straddles. Data from a single stay is not split across training and validation sets, and the inputs are formatted as one-hot vectors. Furthermore, we note that predictions in the test set are only possible for users from the first week with activity in the second week. We use the neural network architecture shown in Figure 3. This architecture maps one-hot representations of users and venues into 12-dimensional embeddings and keeps the day and hour as one-hot. We use 5 hidden layers of 32, 16, 8 and 4 neurons, with Leaky-RELU activations; as well as a binary cross-entropy loss function on the outputted probability of u visiting i at time t. These hyper-parameters were fine-tuned using five-fold cross-validation. We implemented batch-normalization and early stopping based on validation set ROC AUC score.

2.3 Negative Sampling

Most recommender systems suffer from the "single-class" problem and it has been established that their performance depends critically on the methodology to augment the data with negative samples [15, 16]. In our particular problem, we also lack negative data on user locations. One of the contributions of our work is to measure the impact of a negative sampling methodology in the problem of predicting locations using spatial data. A baseline negative sampler which is widely used in recommender systems is random negative sampling (RNS) [17], which samples from the unobserved user-item pairs uniformly at random; typically maintaining a $K : 1$ proportion of negatives to positives for each user.

More sophisticated negative samplers can make use of side information [18], knowledge graphs describing items [15], adversarial generation of negatives [19], among others. These methods perform well because they produce negative samples that are *informative* for the training of the recommender system—likely because they are near the support of the distribution of positive samples and, thus, have non-negligible gradients [15, 20]. However, negative samples should also strive to be *factual* and minimize the risk of false negatives. Several heuristics have been proposed to address the issue of false negatives (e.g. viewed but not clicked items for online purchases [18]) or making assumptions based on the popularity of items not observed (popular items not chosen are likely not preferred [15]). However, this issue remains unresolved for this class of problems and it is expected that a negative sampler producing negatives close to the support of the positives' distribution will significantly

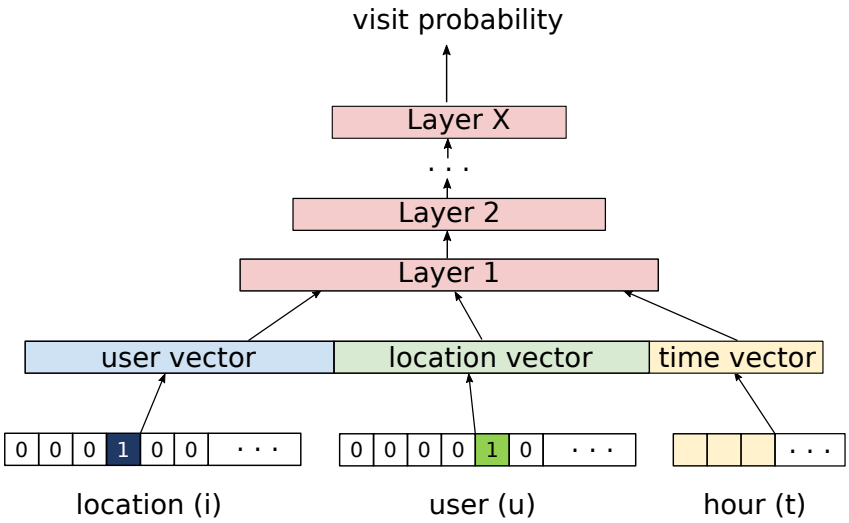

Figure 3: Proposed neural network architecture for inference of missing visits using neural collaborative filtering. users, locations and times (at which users were in the locations) are embedded into a latent space. A nonlinear function predicts the likelihood of a visit from the user to the location or, analogously, of a link in the temporal bipartite graph.

increase the risk of false negatives and affect performance. This is demonstrated in [15], where popularity-based negative sampling (PNS) is outperformed by the benchmark RNS.

An important contribution of our work is our proposed approach to produce high quality negative samples which is specific to this type of data. Namely, we leverage the fact that users are expected to be in only one location during a given time—constraining the temporal user-location bipartite graph. We sample true negatives in the following way: for each user u select an hour t uniformly at random out of the hours for which user u has observed activity (i.e. a link), then, select a location ℓ at random to form the negative $(u,\, t,\, \ell)$. Our benchmark sampler RNS selects ℓ uniformly at random from all the possible locations. To illustrate the benefits derived from restricting the sampler to true negatives, we compare the benchmark with a popularity-based true negative sampler (PTNS), which samples the location ℓ based on popularity. We regularize the frequency distribution as in [21]. For our evaluation on held-out data we consider two metrics:

- **HR@K** or hit-rate at K. Is constructed by predicting probabilities of links of the form $(u,\, t,\, *)$ for every observed link $(u,\, t,\, \ell)$ in the held-out set and obtaining a 1 if the true link is in the top 10 scoring candidates. The HR@K measures the average for all observed links.
- **F1-score** of the binary classification task of predicting whether a link occurs or not. We compute this augmenting the test data with true negatives sampled using RNS in a proportion 10:1.

3 Experiments

3.1 Held-out set validation

In this section, we validate the ability of our model to predict human mobility, with the traditional area under the ROC curve and F1-score metrics. Our finely-tuned model, using PTNS for negative sampling, is tested in the held-out week of (04/08/2020-04/14/2020). We remark that this is held-out set is not ideal because mobility patterns might change from week to week—due both to changes in individuals' circumstances, and also broad external shocks (e.g. government mandates)—and could result in the test set to have a different data distribution. In spite of these limitations, our model is able to accurately predict the mobility patterns of agents in the held-out set. Figure 4 shows the ROC curve, as well as a confusion matrix for a threshold of 0.5. This network obtained an AUC score of 0.7685 and an F1 score of 0.6795 (threshold of 0.5), showing moderate performance in predicting

	all users		high cover users		low cover users	
	HR@10	F1	HR@10	F1	HR@10	F1
RNS	-	0.575	-	0.581	-	0.581
PTNS	-	0.67	-	0.72	-	0.57

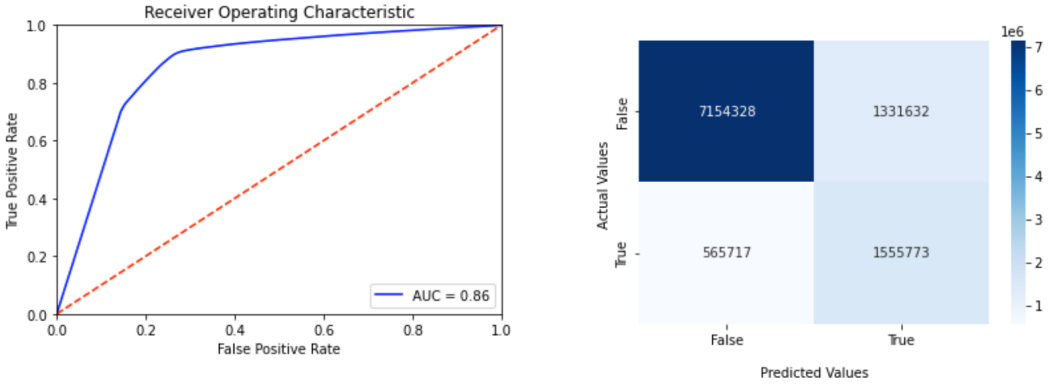

Figure 4: Performance metrics out-of-sample for the fine-tuned NN trained on data from 04/01/2020 - 04/07/2020 and tested on data from 04/08/2020-04/14/2020.

locations out of sample. Our PTNS method outperforms the benchmark for high and low-coverage users.

3.2 User embeddings and stability

Inspired by NLP applications of semantic embeddings [21], we study the low-dimensional representation of users in the embedding layer of our model. Upon the completion of training, the cosine similarity between every pair of users' embeddings is a succinct and efficient metric of the similarity in their mobility patterns. In Figure 5 we can see a T-SNE [22] visualization of the user embeddings after training; a simple K-means clustering has been applied to identify sub-populations with similar mobility patterns, and the weights across sub-populations can be used to construct an epidemic network.

In the absence of ground truth data on which users actually have similar visitation patterns it is difficult to evaluate the quality of these user embeddings. A weaker form of evaluation is to evaluate whether the embeddings of users are stable [23] from one week to the next. We do this by training a model on the two week's worth of data and splitting a fraction of users into two "doppelganger" users, each containing data from one week. We restrict this validation experiment to users with the top 10% of activity since they have a richer set of venues to validate against. Ideally, we would split the data of one user at a time, to test the user's stability in isolation. Because this is impractical, we split the users in batches of 10% at a time. The embeddings derived from training with two weeks of data should result in doppelgangers boasting significantly higher similarity scores than two arbitrary users. Our validation confirms this with doppelgangers having an average similarity score of 0.2366 and an average similarity score between any two users of 0.0065. Figure 5 displays comparisons between the similarity scores of our 10 batches of doppelgangers. When comparing doppelgangers' similarity scores to the distribution of all pairs' similarity scores, the doppelgangers' similarities have an average z-score of 2.8261. In addition, 36.13% of doppelgangers appeared as one of the top 10% most similar users to themselves in the previous week. Taken together, these suggest that doppelganger pairs are embedded more closely together than non-doppelganger pairs are, implying that the neural network indeed learns high-quality embeddings for users.

3.3 Defining contact between users

Although embedding similarity could be a good indicator of users having similar mobility patterns and, thus, having a probability of contact, our framework allows us to directly estimate contact by using the outputs of our trained model. Explicitly, by inputting all combinations of users, venues, and

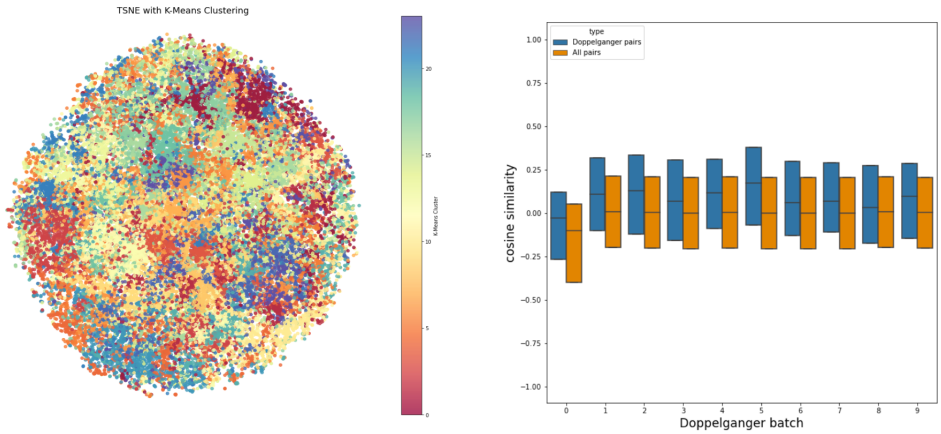

Figure 5: a) TSNE visualization of user embeddings derived from our NCF framework. b) Comparison of cosine similarity between doppelganger pairs and all pairs. The experiment is repeated in ten batches, all but one reflect that doppelgangers are much closer than average, validating the quality of these embeddings.

times, we can estimate the probability that a user u visits venue v at time t. Upon calculating these probabilities, we can estimate the probability that two users are in the same venue in the same hour by summing the probability that 2 users u_1 and u_2 are both at venue v at time t over all venues v. Summing the probabilities of co-location over the 24 hours in a day yields the expected number of hours that two users are at the same location.

4 Conclusion

Since the first results in networked mathematical epidemiology, the central object of study—an epidemic network describing contact between individuals—has remained elusive and difficult to estimate from data. Can GPS human mobility data finally help bridge the gap between theory and applications by providing such networks? In this paper, we propose a novel methodology to construct those contact networks and addresses the main challenge of user-level sparsity in GPS signals. Current contact networks do not factor sparsity into their models, which could lead to problematic inferences of users' mobility trends. Our NCF model avoids learning an incomplete set of information by leveraging similar users' information. Thus, even though individual users do not have complete coverage, NCF can fill in "missing visits" using similar users' behaviors. Our two methods of validation suggest that our model indeed learns users' behavior and can predict future mobility trends at an hourly granularity. This further implies that our model can account for users' dynamic movements and use these fine-grained observations for studying human contact and mobility.

References

[1] J. Keegan and A. Ng, "There's a multibillion-dollar market for your phone's location data," *The Markup*. [Online]. Available: https://themarkup.org/privacy/2021/09/30/theres-a-multibillion-dollar-market-for-your-phones-location-data

[2] E. Pepe, P. Bajardi, L. Gauvin, F. Privitera, B. Lake, C. Cattuto, and M. Tizzoni, "Covid-19 outbreak response, a dataset to assess mobility changes in italy following national lockdown," *Scientific data*, vol. 7, no. 1, pp. 1–7, 2020.

[3] S. Chang, E. Pierson, P. W. Koh, J. Gerardin, B. Redbird, D. Grusky, and J. Leskovec, "Mobility network models of covid-19 explain inequities and inform reopening," *Nature*, vol. 589, no. 7840, pp. 82–87, 2021.

[4] A. Aleta, D. Martin-Corral, A. P. y Piontti, M. Ajelli, M. Litvinova, M. Chinazzi, N. E. Dean, M. E. Halloran, I. M. Longini Jr, S. Merler *et al.*, "Modelling the impact of testing, contact

tracing and household quarantine on second waves of covid-19," *Nature Human Behaviour*, vol. 4, no. 9, pp. 964–971, 2020.

[5] F. Barreras, M. Hayhoe, H. Hassani, and V. M. Preciado, "Autoekf: Scalable system identification for covid-19 forecasting from large-scale gps data," *arXiv preprint arXiv:2106.14357*, 2021.

[6] J. R. Birge, O. Candogan, and Y. Feng, "Controlling epidemic spread: Reducing economic losses with targeted closures," *University of Chicago, Becker Friedman Institute for Economics Working Paper*, no. 2020-57, 2020.

[7] C. C. Kerr, R. M. Stuart, D. Mistry, R. G. Abeysuriya, K. Rosenfeld, G. R. Hart, R. C. Núñez, J. A. Cohen, P. Selvaraj, B. Hagedorn *et al.*, "Covasim: an agent-based model of covid-19 dynamics and interventions," *PLOS Computational Biology*, vol. 17, no. 7, p. e1009149, 2021.

[8] X. He, L. Liao, H. Zhang, L. Nie, X. Hu, and T.-S. Chua, "Neural collaborative filtering," in *Proceedings of the 26th international conference on world wide web*, 2017, pp. 173–182.

[9] P. Covington, J. Adams, and E. Sargin, "Deep neural networks for youtube recommendations," in *Proceedings of the 10th ACM conference on recommender systems*, 2016, pp. 191–198.

[10] L. Minati, M. Frasca, P. Oświcimka, L. Faes, and S. Drożdż, "Atypical transistor-based chaotic oscillators: Design, realization, and diversity," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 27, no. 7, p. 073113, 2017.

[11] G. M. Morton, "A computer oriented geodetic data base and a new technique in file sequencing," 1966.

[12] M. Luca, G. Barlacchi, B. Lepri, and L. Pappalardo, "A survey on deep learning for human mobility," *ACM Computing Surveys (CSUR)*, vol. 55, no. 1, pp. 1–44, 2021.

[13] M. Ester, H.-P. Kriegel, J. Sander, X. Xu *et al.*, "A density-based algorithm for discovering clusters in large spatial databases with noise." in *kdd*, vol. 96, no. 34, 1996, pp. 226–231.

[14] Z. Yang, B. Wu, K. Zheng, X. Wang, and L. Lei, "A survey of collaborative filtering-based recommender systems for mobile internet applications," *IEEE Access*, vol. 4, pp. 3273–3287, 2016.

[15] X. Wang, Y. Xu, X. He, Y. Cao, M. Wang, and T.-S. Chua, "Reinforced negative sampling over knowledge graph for recommendation," in *Proceedings of the web conference 2020*, 2020, pp. 99–109.

[16] X. He, H. Zhang, M.-Y. Kan, and T.-S. Chua, "Fast matrix factorization for online recommendation with implicit feedback," in *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, 2016, pp. 549–558.

[17] S. Rendle, C. Freudenthaler, Z. Gantner, and L. B. Schmidt-Thieme, "Bayesian personalized ranking from implicit feedback," in *Proc. of Uncertainty in Artificial Intelligence*, 2014, pp. 452–461.

[18] J. Ding, F. Feng, X. He, G. Yu, Y. Li, and D. Jin, "An improved sampler for bayesian personalized ranking by leveraging view data," in *Companion Proceedings of the The Web Conference 2018*, 2018, pp. 13–14.

[19] D. H. Park and Y. Chang, "Adversarial sampling and training for semi-supervised information retrieval," in *The World Wide Web Conference*, 2019, pp. 1443–1453.

[20] C. Zhang and C. Li, "Neural collaborative filtering recommendation algorithm based on popularity feature," in *2021 International Conference on Culture-oriented Science & Technology (ICCST)*. IEEE, 2021, pp. 316–323.

[21] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[22] M. LJPvd and G. Hinton, "Visualizing high-dimensional data using t-sne," *J Mach Learn Res*, vol. 9, no. 2579-2605, p. 9, 2008.

[23] L. Wendlandt, J. K. Kummerfeld, and R. Mihalcea, "Factors influencing the surprising instability of word embeddings," *arXiv preprint arXiv:1804.09692*, 2018.