DYNAMIC RECONSTRUCTION OF INFLUENCE GRAPHS WITH ADAPTIVE DIRECTED INFORMATION

B. Oselio, A. Hero

Department of Electrical Engineering and Computer Science
University of Michigan
Ann Arbor, USA

ABSTRACT

We introduce an adaptive version of directed information to estimate an influence graph over nodes with time-varying features. Originally developed as a generalization of the Shannon Mutual Information for quantifying the effect of feedback in a simple communication channel, directed information (DI) measures the amount of causal, time-varying influence that one node's actions have on another node. By estimating these quantities, we can infer a directed graph that captures the flow of influence between nodes. We introduce an online time-averaged version of DI called adaptive directed information (ADI) to study the difference in graphical structure over time. This method is applied to two Twitter US political datasets to track changes in the graphical structure between candidates' Twitter feeds.

***Index Terms*—** Directed Information, Social Networks, Graph Estimation

1. INTRODUCTION

Estimating structure and interaction among targets of interest is a common problem investigated by the signal processing community. Here we are interested in estimating graphical structure that captures directed interactions from observational data generated by multiple agents. Often, it is possible to capture information on the joint behavior of these agents over time. For instance, we may want to infer the interaction of equities in the stock market over time from reported trading activity, or infer social interaction of moving objects in a scene from video. We introduce an adaptive version of the information theoretic measure directed information to quantify these interactions in an on-line recursive fashion.

Directed information (DI) was introduced in [1] to address the problem of feedback in a simple channel. DI can be thought of as an extension of mutual information (MI), and it has extensions to both infinite alphabet channels and continuous time [2]. Graphs created from DI, often called influence graphs, have been explored in the literature previously [3, 4, 5]. The authors of [4] considered influence graph estimation using the well known Granger causality measure that is equivalent to the DI under a Gaussian assumption. The difficulty with DI is that its high computational and sample complexity do not allow for easy and scalable estimation methods, especially when the data is high dimensional, non-Gaussian and discrete. We describe a method that allows a time-varying DI graph to be estimated under a Markov model.

Standard DI, while able to take into account time-varying properties of the agents over time, is insensitive to abrupt changes in interaction, dependence, and influence, due to its heavy weight on past observations. We introduce adaptive directed information (ADI) that modifies DI so that it is more capable of picking up subtle shifts in the nodes' interactions. We show in this paper that, under a Markov assumption, the ADI can be computed efficiently in an online fashion using a recursive updating scheme over time. To our knowledge, with the exception of our preliminary work [6] and the fuller treatment given in this paper, the recursive update form of the ADI has not appeared elsewhere in the literature. Under a simplifying instantaneous conditional independence assumption the ADI updates depend only on the joint distributions of third order. To illustrate the ADI we apply it to two Twitter datasets to estimate the influence graph among Twitter users.

This paper is organized as follows: Section 2 discusses related work on influence estimation, DI, and DI graphs. Section 3 will introduce the problem and some notation conventions. 4 will introduce the concept of DI and ADI. Section 5 will demonstrate the chosen model for text information, and some assumptions made to make ADI estimation tractable. Section 6 will explain the process of generating DI and ADI graphs. Section 7 will introduce the two Twitter datasets, and discusses results from the described methods. Finally, Section 8 summarizes the contributions of the paper.

2. RELATED WORK

Influence among actors has been studied in many settings [7, 8, 9]. DI has been studied extensively both theoretically and in the context of applications. The estimation of the di-

This research was partially funded by ARO Grants W911NF-12-1-0443 and W911NF-15-1-0479.

rected information rate for stationary ergodic processes has been studied in [10]. Some applications of DI are covered in [11, 12] regarding gambling and portfolio theory. In addition, [13] uses DI to infer biological regulatory networks.

DI graphs have also been studied, most recently in [3], which focuses on the estimation of the causal DI graph, as well as DI estimation. The authors of [3] identify sample complexity for both non-parametric and parametric estimators for DI. The focus of [3] is on cases where the processes are stationary. [4] also discusses DI graphs, with their focus on the relationship to Granger causality. To our knowledge, no other group has introduced an adaptive version of DI.

3. SETUP AND NOTATION

Consider a set of n agents $(N_1, N_2, \dots N_n)$, represented as nodes in a graph, that generate P-dimensional features that evolve over T time samples. We assume that the features are binary. We denote a random vector evolving over a time period t as a capital letter with a subscript, e.g., X_t. A capital letter with a superscript T represents the random vectors up to and including T, $X^T = X_1, X_2, \dots, X_T$. Finally, a lowercase letter with a superscript and a subscript, x_t^i, represents the scalar random feature i at time t.

4. DIRECTED INFORMATION

4.1. Definition and Properties

Directed information is an information theoretic measure originally introduced by [1] to study the effect of feedback on channel capacity. Given a discrete communications channel $P(Y_t|X^t, Y^{t-1})$, with input time series $X_1, X_2, \dots, X_t$ and outputs $Y_1, Y_2, \dots, Y_t$, the directed information (DI) is defined as:

$$DI(X^T \to Y^T) = \sum_{t=1}^{T} I(X^t; Y_t|Y^{t-1}). \tag{1}$$

The DI is asymmetric, $DI(X^T \to Y^T) \neq DI(Y^T \to X^T)$. Furthermore, when the channel exhibits no feedback, e.g.,

$$P(X_t|X^{t-1}Y^{t-1}) = P(X_t|X^{t-1}), \tag{2}$$

DI is equivalent to the standard Shannon mutual information [14].

4.2. Adaptive Directed Information

DI can account for the time-varying nature of interaction among targets (i.e. changing $P(Y_t|X^tY^{t-1})$), but does not vary over time and places equal weight on each time point in the time series. We introduce the adaptive directed information (ADI) as a time varying modification of DI defined as a discrete time filter $g(t,i)$ applied to the sequence $I(X^i; Y_i|Y^{i-1})$, $i = 1, \dots, \infty$:

$$(ADI_{N_x \to N_y})_t = \sum_{i=1}^{t} g(t,i) I(X^i; Y_i|Y^{i-1}) \tag{3}$$

The filter, $g(t,i)$ can be chosen in various ways, including the windowed exponential $g(t,i) = e^{-(t-i)\lambda}c_t, i \leq t, \lambda > 0$, where $c_t = (1-e^{-\lambda})/(1-e^{-(t+1)\lambda})$, or the uniform window of length T, $g(t,i) = 1/T, |t-i| \leq T$.

5. EMPIRICAL ESTIMATION OF DI AND ADI

Empirical estimation of the DI and ADI from data poses challenges, especially in high feature dimension P. The complexity of estimation can be reduced by imposing Markov assumptions, performing dimension reduction on the feature space, and making simplifying approximations to the joint distributions. Under a jointly Markov assumption on the pair of time series $\{(X_i, Y_i)\}_i$ we obtain a simplification of the following conditional probabilities:

$$P(X_t, Y_t|X^{t-1}, Y^{t-1}) = P(X_t, Y_t|X_{t-1}, Y_{t-1}), \tag{4}$$

$$P(X_t|X^{t-1}) = P(X_t|X_{t-1}), \tag{5}$$

$$P(Y_t|Y^{t-1}) = P(Y_t|Y_{t-1}). \tag{6}$$

The Markov representations (4-6) reduce the joint distribution on the left side of (4), which depends on the entire past $\{(X_i, Y_i)\}_{i=1}^{t-1}$, to the right hand side of (4), which only depends on the most recent past (X_{t-1}, Y_{t-1}). Thus (4) can be computed from the joint distribution of the four variables $\{X_t, Y_t, X_{t-1}, Y_{t-1}\}$, what we call a fourth order distribution. One can simplify further by imposing the additional "instantaneous conditional independence" property that X_t and Y_t are independent given past information:

$$P(X_t, Y_t|X_{t-1}, Y_{t-1}) = P(X_t|X_{t-1}, Y_{t-1})P(Y_t|X_{t-1}, Y_{t-1}), \tag{7}$$

which only involves third order distributions. In order to exploit this factorization to estimate DI and ADI, we write DI in terms of conditional entropies:

$$DI(X^T \to Y^T) = \sum_{t=1}^{T} H(Y_t|Y^{t-1}) - H(Y_t|Y^{t-1}, X^t) \tag{8}$$

Using (4-6), we obtain:

$$DI(X^T \to Y^T) = \sum_{t=1}^{T} H(Y_t|Y_{t-1}) - H(Y_t|Y_{t-1}, X_t, X_{t-1}). \tag{9}$$

$$DI(X^T \to Y^T) = DI(X^{T-1} \to Y^{T-1}) + H(Y_T|Y_{T-1}) - H(Y_T|Y_{T-1}, X_T, X_{T-1}). \tag{10}$$

Using standard properties of conditional entropy and (7), the DI expands to

$$\begin{aligned} DI(X^T \to Y^T) &= DI(X^{T-1} \to Y^{T-1}) \\ &\quad - H(Y_{T-1}) - H(Y_T, Y_{T-1}, X_T, X_{T-1}) \qquad (11) \\ &\quad + H(Y_{T-1}, X_T, X_{T-1}) + H(Y_T, Y_{T-1}) \\ &= DI(X^{T-1} \to Y^{T-1}) + H(Y_T, Y_{T-1}) - H(Y_{T-1}) \\ &\quad - H(Y_T|X_{T-1}, Y_{T-1}) - H(X_T|X_{T-1}, Y_{T-1}) \\ &\quad - H(X_{T-1}, Y_{T-1}) + H(Y_{T-1}, X_T, X_{T-1}). \qquad (12) \end{aligned}$$

Hence, the DI can be computed from third order distributions in recursive form where only third order entropy is required for updating the DI at time $T-1$ to time T.

We can calculate ADI directly from DI, but if we choose to use an windowed exponential filter $g(t,i)$, we obtain the recursion:

$$\begin{aligned} (ADI_{N_x \to N_y})_t &= \alpha(ADI_{N_x \to N_y})_{t-1} + (1-\alpha)[H(Y_T, Y_{T-1}) \\ &\quad - H(Y_{T-1}) - H(Y_T|X_{T-1}, Y_{T-1}) - H(X_T|X_{T-1}, Y_{T-1}) \\ &\quad - H(X_{T-1}, Y_{T-1}) + H(Y_{T-1}, X_T, X_{T-1})], \qquad (13) \end{aligned}$$

where $\alpha = (e^{-\lambda} - e^{-(t+1)\lambda})/(1 - e^{-(t+1)\lambda})$.

5.1. Estimating Joint Distributions of Binary Vectors

Under the instantaneous conditional independence assumption the third order distributions of the form $P(X_T, Y_T, Y_{T-1})$ must be estimated in order to calculate ADI. We implement this estimator by binning together groups of time samples in order to estimate the distributions.

For concreteness we specialize to feature vectors $X = [x^1, \ldots, x^P]$ and $Y = [y^1, \ldots, y^P]$ with binary elements, i.e., $x^i, y^i \in \{0, 1\}$. While any feature dependency model could be accommodated, for simplicity we will assume elementwise independence of the feature vectors — namely, that the j-th scalar feature x_j^t is jointly independent of the other scalar features x_i^t and y_i^t, for $i \neq j, t = 1, \ldots, T$. This allows us to factorize the joint distributions of three feature vectors into third order distributions of scalar variables. Hence, for example,

$$\begin{aligned} P(X_n, X_{n-1}, Y_{n-1}) &= \prod_{i=1}^{P} P_i(x_n^i, x_{n-1}^i, y_{n-1}^i) \qquad (14) \\ &= \prod_{p=1}^{P} \theta_{p_1}^{(1-t_1)(1-t_2)(1-t_3)} \theta_{p_2}^{(1-t_1)(1-t_2)(t_3)} \cdots \theta_{p_8}^{t_1 t_2 t_3}. \qquad (15) \end{aligned}$$

$\{\theta_{p_i}\}$ are parameters that must be estimated. We propose using maximum likelihood estimators with Stein regularization [15]:

$$\hat{\theta}_{p_i} = (1 - \lambda_S)\hat{\theta}_{p_i}^{ML} + \lambda_S, \qquad (16)$$

where $\hat{\theta}_{p_i}^{ML}$ is the maximum likelihood estimate of $\theta_{p_i}^{ML}$, and λ_S can be chosen to optimize bias-variance tradeoff as in [15].

The factorization (17) allows the entropy to be computed from individual feature entries:

$$H(Y_{T-1}, X_T, X_{T-1}) = \sum_{i=1}^{P} H(y_{T-1}^i, x_T^i, x_{T-1}^i). \qquad (17)$$

We will apply the proposed ADI estimator to text data, specifically corresponding to the content of tweets from Twitter. From this data, we bin the tweets, forming documents of collected tweets over time, and model each word as a binary random variable indicating its presence or absence. These vectors are then used to estimate the $\{\theta_{p_i}\}$ parameters.

5.2. Computational and Model Complexity

Each probability estimate for a third order distribution takes $\mathcal{O}(t)$ computations, where t is the number of samples used to calculate the estimate. There are $\mathcal{O}(P)$ entropies to calculate for each estimate of directed information, and each entropy can be calculated in $\mathcal{O}(1)$. We must calculate the DI T/t times for each pair, and there are $n(n-1)/2 = \mathcal{O}(n^2)$ pairs. In total, calculation of every pairwise DI in the graph requires $\mathcal{O}(TPn^2)$ computations. ADI has an identical complexity analysis. For each DI calculation, we estimate $16P$ parameters, and these parameters can be used for both orderings of the pair. Therefore, our method must estimate $(16PTn(n-1))/(2t)$ parameters. This compares favorably with other methods that attempt to estimate higher order distributions; for general vectors of binary features and pairwise DI, one must estimate $\mathcal{O}(2^P)$ parameters for each pair.

6. CREATING INFLUENCE NETWORKS

Once pairwise DI and ADI have been calculated for all n nodes, we are able to infer graphical structure. The most naïve way to do this is to simply use each non-zero DI entry as a directed weighted edge between targets; this approach can be quite noisy. A more reasonable approach is to create a hypothesis test for each edge, and only keep the edges that have a statistically significant influence.

For DI, there are two possible ways to do this. One method, the approach of [16], uses a functional transformation leading to approximation of p-values for existence of an edge. Another method, proposed in [15], invokes a central limit theorem for DI. In this paper, the latter approach is used.

7. APPLICATION TO TWITTER DATASETS

The methods described above are applied to two datasets. The first, which is a dataset regarding the United States Presidential primary candidates, are all the tweets from the campaign Twitter accounts of each candidate from Oct. 1st, 2015 to

Jan. 13th, 2016. The second dataset is of the members of the United States Senate, over the same time period.

7.1. 2015 US Presidential Candidates Dataset

This dataset consists of 15 primary candidates. In total, there are 8918 tweets in the dataset. After cleaning and stemming, and binning the tweets into 12-hour time periods, the features (words) are further filtered as follows: if the word is used in less than 10 of the bins or greater than 50% of them, it is discarded. In total, 1554 features remain.

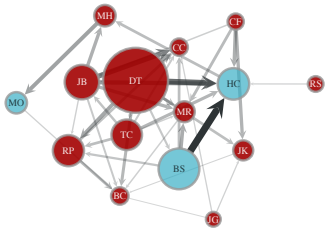

Fig. 1. Relative DI network of US Presidential primary candidates. The width of the directed edge as well as the shade is related to the magnitude of the DI, and the size of each node represents the volume of tweets.

Fig. 1 shows the relative DI for the entire time period, after hypothesis testing at a 5% family-wise error rate probability, where the magnitude of relative DI is $|DI_{X^T \to Y^T} - DI_{Y^T \to X^T}|$, and the direction of the arrow represents the sign (arrow points towards N_y if $DI_{X^T \to Y^T}$ is larger). The width and shade of the directed edge is related to the magnitude of the relative DI. Further, the size of each node represents the volume of tweets. The network in Fig. 1 has some interesting properties. First, we see that nodes such as Hillary Clinton and Rand Paul are sinks of influence, that is they have high indegree and are influenced by many others. Conversely, there are nodes with high outdegree, such as Jeb Bush and Bernie Sanders that are less influenced by others.

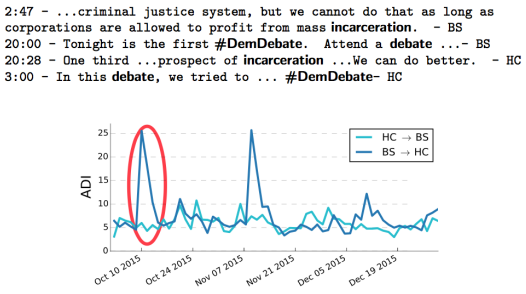

Fig. 2. ADI for Bernie Sanders and Hillary Clinton. Above the graph are representative tweets related to the circled spike.

Fig. 2 demonstrates the utility of ADI. ADI was calculated using an windowed exponential filter with $\lambda = 0.7$. Using ADI, we are able to see the time-varying nature of influence, this time specifically between Bernie Sanders and Hillary Clinton. We see two large spikes in the ADI over time. The tweets above the graph partially contribute to the circled spike. Specifically, we see that Bernie Sanders was discussing incarceration and the upcoming Democratic debate before Hillary Clinton does, which results in a spike of ADI from Bernie Sanders to Hillary Clinton.

7.2. 2015 US Senatorial Dataset

The Senatorial dataset consists of tweets from 80 of 100 US senators over the period Oct. 1st, 2015 to Jan. 13th, 2016. The remaining 20 senators were excluded due to lack of tweet volume. In total, the dataset consists of 96090 tweets. Using a similar process as the Presidential candidates dataset, after cleaning, stemming, and binning, we obtained 1230 features. For the Senatorial dataset we took the bin time length to be 1 day, and took 3 bins over each estimation of DI.

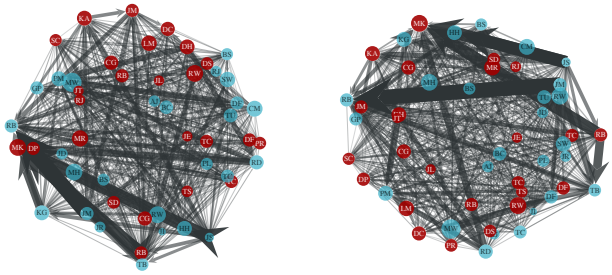

Fig. 3. ADI network of US Senators over two consecutive time periods from left to right.

Fig. 3 are two relative ADI networks of consecutive time periods of the senators studied. ADI was calculated using an windowed exponential filter with exponential parameter $\lambda = 0.7$. These edges were chosen by hypothesis testing with a 5% family-wise error rate probability. Some senators are not displayed as they have no significant edges. We notice that there are nodes of high activity such as RB (Rob Bishop) and MK (Marcy Kaptur). Further, we see significant evolution in the network, with nodes adapting their behavior; this shows the method's ability to estimate changes in influence.

8. CONCLUSION

We presented an adaptive version of directed information, called ADI. ADI better captures time-varying interactions between agents in a network by representing the time evolution of DI as the output of a discrete filter with instantaneous DI as input. We further presented efficient, recursive methods to compute DI and ADI under Markovian and conditional independence assumptions. Finally, we illustrated these methods on two political Twitter datasets from the 2015 US Presidential campaign.

9. REFERENCES

[1] J Massey, "Causality, feedback and directed information," *Proc. Int. Symp. Inf. Theory Applic.*, pp. 303–305, 1990.

[2] Tsachy Weissman, Young-Han Kim, and Haim H. Permuter, "Directed Information, Causal Estimation, and Communication in Continuous Time," *IEEE Transactions on Information Theory*, vol. 59, no. 3, pp. 1271–1287, 2012.

[3] Christopher Quinn, Negar Kiyavash, and Todd P. Coleman, "Directed Information Graphs," *IEEE Transactions on Information Theory*, vol. 61, no. 12, pp. 6887–6909, 2015.

[4] Pierre-Olivier Amblard and Olivier J. J. Michel, "On directed information theory and Granger causality graphs," *Journal of Computational Neuroscience*, vol. 30, no. 1, pp. 7–16, 2011.

[5] Jalal Etesami, Negar Kiyavash, and Todd P. Coleman, "Learning minimal latent directed information trees," *IEEE International Symposium on Information Theory - Proceedings*, pp. 2726–2730, 2012.

[6] Brandon Oselio and Alfred Hero, "Dynamic Directed Influence Networks : A Study of Campaigns on Twitter," in *Social, Cultural, and Behavioral Modeling, 9th International Conference*, 2016, pp. 152–161.

[7] Manuel Gomez-Rodriguez, Jure Leskovec, and Andreas Krause, "Inferring networks of diffusion and influence," *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '10*, vol. 5, no. 4, pp. 1019–1028, 2010.

[8] Brian Baingana, Gonzalo Mateos, and Georgios B. Giannakis, "Proximal-gradient algorithms for tracking cascades over social networks," *IEEE Journal on Selected Topics in Signal Processing*, vol. 8, no. 4, pp. 563–575, 2014.

[9] Maryam Tahani, Afshin Hemmatyar, and Hamid R. Rabiee, "Inferring Dynamic Diffusion Networks in Online Media," *ACM Transactions on Knowledge Discovery from Data*, vol. 10, no. 4, pp. 44, 2016.

[10] Jiantao Jiao, Haim H. Permuter, Lei Zhao, Young-Han Kim, and Tsachy Weissman, "Universal Estimation of Directed Information," *IEEE Transactions on Information Theory*, vol. 59, no. 10, pp. 6220–6242, 2013.

[11] Haim H. Permuter, Young Han Kim, and Tsachy Weissman, "On directed information and gambling," in *IEEE International Symposium on Information Theory - Proceedings*, 2008, pp. 1403–1407.

[12] Haim H. Permuter, Young-Han Kim, and Tsachy Weissman, "Interpretations of Directed Information in Portfolio Theory, Data Compression, and Hypothesis Testing," *IEEE Transactions on Information Theory*, vol. 57, no. 6, pp. 3248–3259, 2011.

[13] Arvind Rao, Alfred O. Hero, David J. States, and James Douglas Engel, "Using directed information for influence discovery in interconnected dynamical systems," *Proc. SPIE 7074, Advanced Signal Processing Algorithms, Architectures, and Implementations XVIII, 70740P*, 2008.

[14] Thomas M. Cover and Joy a. Thomas, *Elements of Information Theory*, 2005.

[15] Xu Chen, Zeeshan Syed, and Alfred Hero, "EEG spatial decoding with shrinkage optimized directed information assessment," *ICASSP 2012 Proceedings*, pp. 577–580, 2012.

[16] Arvind Rao, Alfred O Hero, David J States, and James Douglas Engel, "Using directed information to build biologically relevant influence networks.," *Journal of bioinformatics and computational biology*, vol. 6, no. 3, pp. 493–519, 2008.