Networks of Diffusion and Centers of Scribal Innovation in Classic Maya Society

Abstract

Classic Maya hieroglyphic writing is one of the best-documented and thoroughly deciphered scripts in prehispanic Mesoamerica, yet we still know remarkably little about how this writing system evolved. Specifically, when and how did novel graphemes – the discrete and most basic units of text – spread across the Maya region? A major challenge to inferring influence is that who–influence–whom network is largely an intangible and unobservable phenomenon. Moreover, the underlying network of interactions through which influence potentially propagates is, at best, partially observable in most cases. Taking into consideration these challenges, we devise a probabilistic framework for building networks of grapheme innovation using the observed dates of grapheme inscription and the geographic positions of archaeological sites. We use the Susceptible-Infected (SI) transmission model from mathematical epidemiology as the basis of our proposed method. The inferred influence networks of grapheme innovation are validated by comparing them with the sociopolitical ties documented in the written texts. These influence networks facilitate the identification of sites that stood out as innovators at various points during the Classic period. The proposed probabilistic model is applicable to a wide range of archaeological network inference problems and opens up further questions about co-evolution, replacements, and the discontinuation of graphemes through time.

1 Overview

In any social or political environment, it is usual that autonomous entities, share, emulate, adopt, and influence one another's tangible and intangible culture. Understanding the mechanism of these adaption and influencing processes has fascinated researchers in a wide range of scientific arenas for decades now. Historians and archaeologists have long hypothesized that the Maya communities were complex networks of interdependent societies that interacted politically as well as socially and impacted each others culture in interesting ways. *"Maya hieroglyphic writing, in particular the tradition of inscribing texts and images on carved stone monuments, offers evidence for widespread and mutually intelligible cultural practices that were, at the same time, neither unchanging nor uniform in their semantic content."* [13]. This work is part of a similar endeavor in deducing to what an extent the syntactic and semantic similarities in the Classic Maya inscriptions could be attributed to influence among the Classic Maya communities, whether it be social or political. We look into the records of Classic Maya hieroglyphic writings to trace the use of each grapheme used in their writings. Given the timestamped inscription from each Maya site, we infer the most likely path of influence by factoring in the temporal information of the inscriptions as well as the proximal information of the sites being compared. To guage how much the inferred paths of influence replicate the social and political structure of the Classic Maya communities we compare the inferred network of influence to the social and political relations among those communities, declared in the text of those inscriptions. Thus, by using two mutually non–overlapping and qualitatively different subsets of the *Maya Hieroglyphic Database*, we attempt to infer networks of influence in geographically proximal yet autonomous Maya communities through shared language structure and usage. Precisely, this work can be divided into two parts:

- Based on the large amount of inscriptions pertaining to the late Classic period found at different Maya sites, linguistics have been able to decipher the structure and meaning of the language used by the Maya communities [**?**]. A remarkable feature of this large quantity of text found at different sites

is the similarity in the language used by geographically disparate sites [**?**]. This artifact raises the question of how much this similarity can be attributed to the mutual sharing and influencing among disparate communities? We build inference models that extract strong influence paths conditioned on temporal information about the inscriptions as well as the proximal information of the sites using similar graphemes.

- The text of the Maya inscriptions allude to political and social relations among the Classic Maya communities. Thus the actual meaning of the text helps understand the nature of social and political relations among those communities. Since these Maya communities engaged with each other through a variety of political, social, and familial ties; it is natural to assume that the similarity exhibited in their use of language could have resulted, to some extent, by influence of those ties. Hence, we can leverage the information about those various forms of relations to verify and validate the strength of the paths of influence deduced by studying the utility of graphemes by those sites.

In the first part of this work we introduce a methods to infer probable paths of spread of graphemes among the Maya communities by leveraging the timestamps of the inscriptions and the proximity of the sites. Further, we use the records about political and familial affiliations, as partially observed ground truth, to measure the accuracy of our suggested approaches.

2 Related work

How information, materials, diseases propagate in a network of social organisms is a challenging question that has occupied researchers in a wide range of disciplines, such as social science, epidemiology, and viral marketing, among others. Depending on the nature of the process, it is usually found that there is a systematic pattern by which the phenomenon spreads over a set of entities. This concept has fascinated researchers in a myriad of fields for decades now. Especially the question of how to approximate the mechanism by which the diffusion occurs for better understanding of orientation and scope of the spread? For example, mathematical epidemiology has long been invested in translating the spread of chronic and infectious diseases into linear or non–linear system and understanding more complex causal relations of the spreading disease [2, 3]. In social and behavioral sciences as well as in marketing, diffusion of information, ideas, and adaption of new products are some of the practical concepts being thoroughly analyzed to maximize the gains of information spread [9]. With the advent of web and relatively easy availability of massive amount of online social networks data, in computer science, the theoretical foundations of many diffusion process as well as their algorithmic and computational complexity has become one of a hotly researched topic [8]. Theoretically how to maximize the diffusion of a spreading phenomenon or to predict the size and direction of the spread is one angle by which this problem is being studied [10, 12]. On the other hand a very natural and genuinely perplexing question is how to infer a diffusion process given we do not know of any well defined mechanism [1, 5, 4, 6, 11]. Also, the availability of large amount of data creates new kind of challenges such as missing data or false positives [11, 14].

In archaeology, the study of social, political, and cultural movements overlap with concepts of diffusion in networks in many cases [**?**]. Even though on the surface it would appear that the same theories about diffusion from social and behavioral sciences could be adapted to historical networks, the data from archaeology has its specific challenges that hinder the direct applications of methods from other fields. Missing and sparse data, subjective interpretation of data, and reliance of proxy data in the stead of actual flow data are some of the constraints that are imperative to take into consideration before analyzing the data. Diffusion related research in archaeolgy. Especially the recent and influential ones.

Classic Maya data .... Ritual Diversity and Divergence of Classic Maya Dynastic Traditions: A Lexical Perspective on Within-Group Cultural Variation [13]. More references to Classic Maya research on social ties and ritual diversity.

The overall goal of this work is to infer network of influence among classic maya tribes through the similarity in their text. These groups of people were known to engage in social, political, and familial relationships as evidenced by the interpretation of their writings [**?**]. However, an open and fascinating

question that still lacks a systematic undertaking is how much these communities influenced each other through their various social, political, and familial roles? For example, through their writings its clearly evidenced that these communities engaged in a number of cultural practices independently [**?**]. However, the presence of these activities in various communities allude to the fact that their must be some level of influence going on among these communities [**?**]. The question then arises is that how well can we understand that influence given all the indirect links among these communities? In this work we use the basic units of their text that were shared in their writings as a proxy to their probable influence on each other. We build on the inference model presented in [5]. This work develop a method to infer paths of diffusion by using the time of adoption of a piece of information by each entity. We extend this model in four key ways: *a*) incorporate the geographical proximity of the entities with the temporal information to make the model more informative, *b*) instead of using only one time stamp as the evidence of adoption, we take into consideration the entire interval in which an entity was "alive", i.e. multiplicity of usage of a textual unit is used to strengthen the amount of influence one site can exert on another, *c*) we relax the single source of influence assumption to build more complex influence graphs instead of a sparse tree, and *d*) instead of aggregating the spread of all the "contagion" into one tree, we analyze each diffusion independently.

To the best of our knowledge there is no systematic work done on interpreting *diffusion phenomenon with multiple time stamps associated with each entity*. However, it is natural that in most real–world dynamic processes there can potentially be more than one time existence of a spreading concept. For example, online blogs can repeatedly be talking about the same topic, an infectious disease can recur in a living being, customers may buy the same item repeatedly, among others. Moreover, in many such diffusing concept an interval of time can be associated with each entity such that the entity is assumed to be "alive" or actively using or practicing the spreading concept. This repeated use creates a kind of visibility of the entity using the spreading concept and could influence the next entity with various degrees of strength depending on the assumptions of the diffusion mechanism. Lastly, since the strength of influence is primarily dependent on the time difference of adoption of two sites, it is logical to as much consider the entire set of times at which the influencing site used the spreading concept to estimate when and how much it could have most likely influenced the later adopting site.

In the next section we explain the application for which we build this inference of influence model.

3 Maya Hieroglyphic database

The Maya Hieroglyphic database project is in an interdisciplinary effort to catalog and analyze the graphic linguistic, and semantic information encoded in a comprehensive sample of ancient Maya texts. The MHD is a unique catalog of hieroglyphic texts that encode comprehensive spatial, temporal, and linguistic information about Classic Maya script and language. Extant database consists of $\sim 75,000$ records of glyph blocks, which are discrete units of text composed of variable combinations of *graphemes* to form a word or phrase. These records represent 247 distinct archaeological sites.

In this work we use two subset of the larger dataset: *a*) time stamped records of use of each grapheme, *b*) social, political, and familial dyadic relations . Both these datasets come from texts inscribed on hieroglyphic monuments from known archaeological sites. These texts also include dates for when the text was written as well as the events described in the text (these are often retrospective in nature).

Following is a brief explanation of what information relevant to the current analysis each of the dataset contains.

3.1 Grapheme dataset

Graphemes are the most basic unit of writing in the Maya hieroglyphic script. About 956 unique graphemes were used during the Classic period. While many signs have been deciphered, this dataset also includes unique graphemes that are not yet deciphered. The common use of these graphemes by distinct communities shows they shared language. Hence, have had mutually influential social relations. Using the dated records of individual grapheme inscriptions by disparate site we infer the likely paths of influence among the sites.

3.2 Relationships dataset

This is the subset of the data that depicts dyadic relationships of six distinct categories among the various Maya sites. These relationships are: *antagonistic, diplomatic, kinship/dynastic, nametag, subordination, and unknown.* All the recorded pairwise relations are asymmetric and dated. We use the existence of any relation as indicative of social contacts between the pair of communities. And thus use this supplementary information as a validation of the inferred path of influence between two communities.

4 Influence propagation pathways

Given the temporal information of inscriptions on disparate site, the question we answer here is that if similarity in language structure is the result of influence among the disparate sites, then how to infer the most likely paths through which communities could have influenced each other. To be able to infer the influence graph, one has to make certain assumptions and rules about the mechanism of influence propagation.

Before describing the model formally, we state the assumptions built into the model that have a direct implication to the paths inferred and consequently on the conclusions drawn from it.

4.1 Assumptions

1. Each grapheme is unchanging at least for the time period of observations under study. They are all independent of each other, that is, for example, occurrence of one does not precludes to the occurrence of another or one grapheme cannot be seen as a transformation of another or the semantics of graphemes do not have any bearing to their co–occurrence. These are of course, simplifying assumptions and can have a non–trivial affect on the resulting influence structure. However, incorporating these constraints would restrict the model considerably such that it would be hard to generalize it to an overall diffusion mechanism.

2. For the influence process it is assumed that the first time a record of a grapheme is found on a site is the time it was "influenced". All subsequent recording of the same grapheme on the same site is used as an indicator for the length of the time that the site was "actively" utilizing that grapheme. See Figure 1 for the timeline of inscriptions of a sample of graphemes. Also, Figure 2 shows the frequency distribution of inscriptions of grapheme over time.

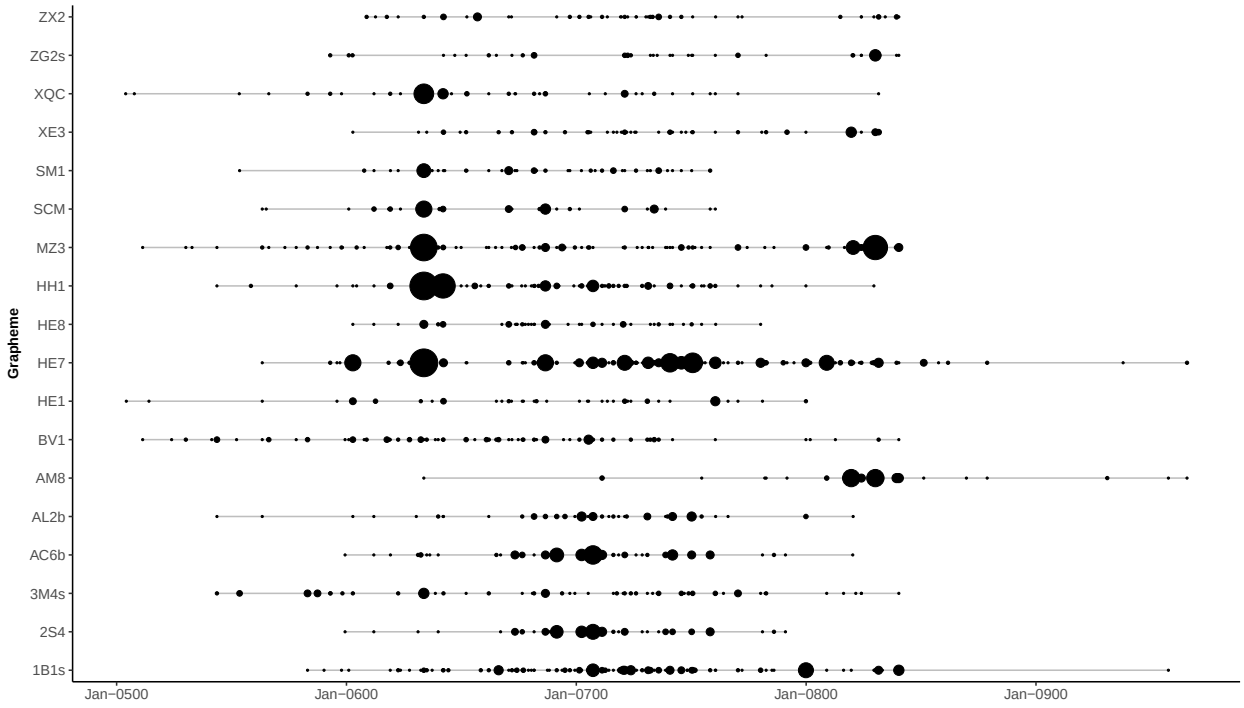
3. The first time of inscription at a site determines the directionality of the influence. That is, a site with earlier first inscription can potentially influence all the sites with the later first inscription time and not vice versa.

4. However, the strength of influence is determined by the latest inscription of a grapheme on a site to the first time of inscription of the same grapheme on another site.

5. A model specific assumption: each site can potentially be influenced by one other site in adopting a grapheme. In network terms, each node can have only one parent. In the path inference model the parent site is chosen by selecting the site with the highest weight of influence leading to the influenced site.

Figures 3,4, and 5 exemplifies the influence process in a simulated scenario. Figure 3 depicts five sites $\{u, v, w, x, y\}$ ordered vertically from the earliest to latest time of first time inscription of a certain grapheme. The horizontal lines across each site indicate the time line of inscription of the grapheme on that site. For example, u inscribed the grapheme at t_1, t_2, t_4, t_7, and t_{10}. Assumption 2 can be visualized by Figure 3. Figure 4 exemplifies the assumption 3. It shows the potential source of influence for each site. Since u is the first site to inscribe the grapheme, it has no influencer. It is the so-called "innovator". It can potentially influence every site subsequently inscribing the same grapheme for the first time. However, how influential one site to another is determined by the difference in the latest inscription of an already inscribing site to the

Figure 1: Timelines of grapheme inscriptions



first time a site inscribes the grapheme (Assumption 4). This is depicted in Figure 3, as only one example, through the blue dotted line between time t_4 (3rd inscription of site u) and t_6 (first inscription of site w). t_4 is the latest time of inscription of the grapheme by u before w adopted it. Hence, the influence of u on w is the function of the time difference between t_4 and t_6. In Figure 3 the rest of the weights of influence edges are omitted to avoid clutter. Once the *Directed Acyclic Graph (DAG)* for each graph is built, the best (strongest) source of influence for each site is picked. This is the assumption 5 which is depicted in Figure 5.

4.2 Framework of inference of influence propagation pathways

In this section we formally explain the model of inference of propagation of influence. Table 1 lists all the notations used to describe the method formally.

4.2.1 Influence propagation model

The overall mechanism of the influence propagation model is based on the Independent Cascade Model [10]. The model works as follows. Sites are in one of the two states: influenced or susceptible. Each influenced site can influence each susceptible site. Once a site is influenced it remains in that state indefinitely. Each influence is independent of all others, that is, each site is influenced by every already influenced site independently.

Using this propagation model, the *inference of influence pathways* takes place through the following three steps of refinement.

Influence propagation DAG Since we do not have an underlying graph to simulate the diffusion process, we first build the baseline graph over which the propagation can logically happen. This graph is the *Directed*

Figure 2: Frequency of inscriptions of graphemes over time

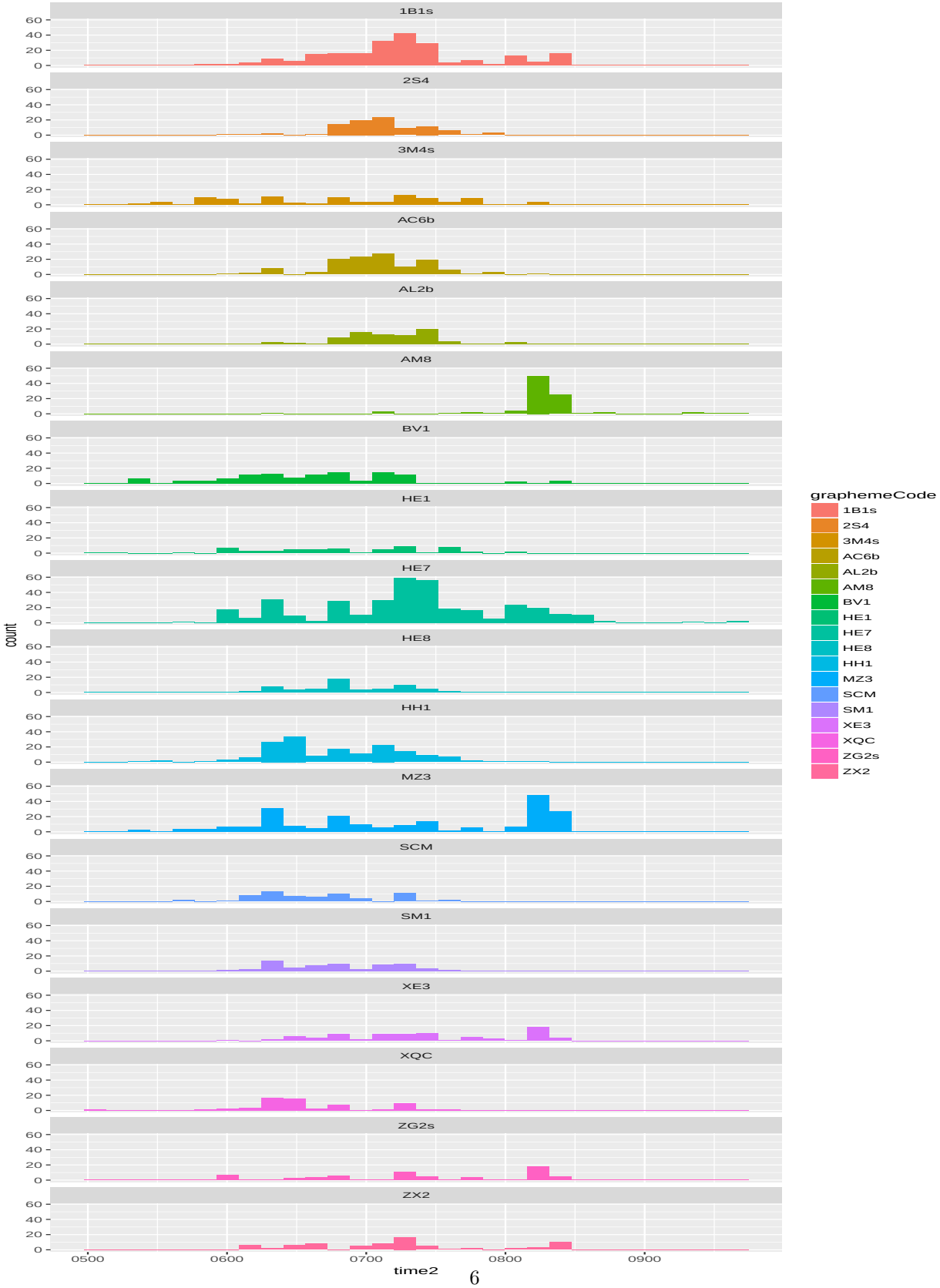


Figure 3: Timeline of inscriptions of a certain grapheme

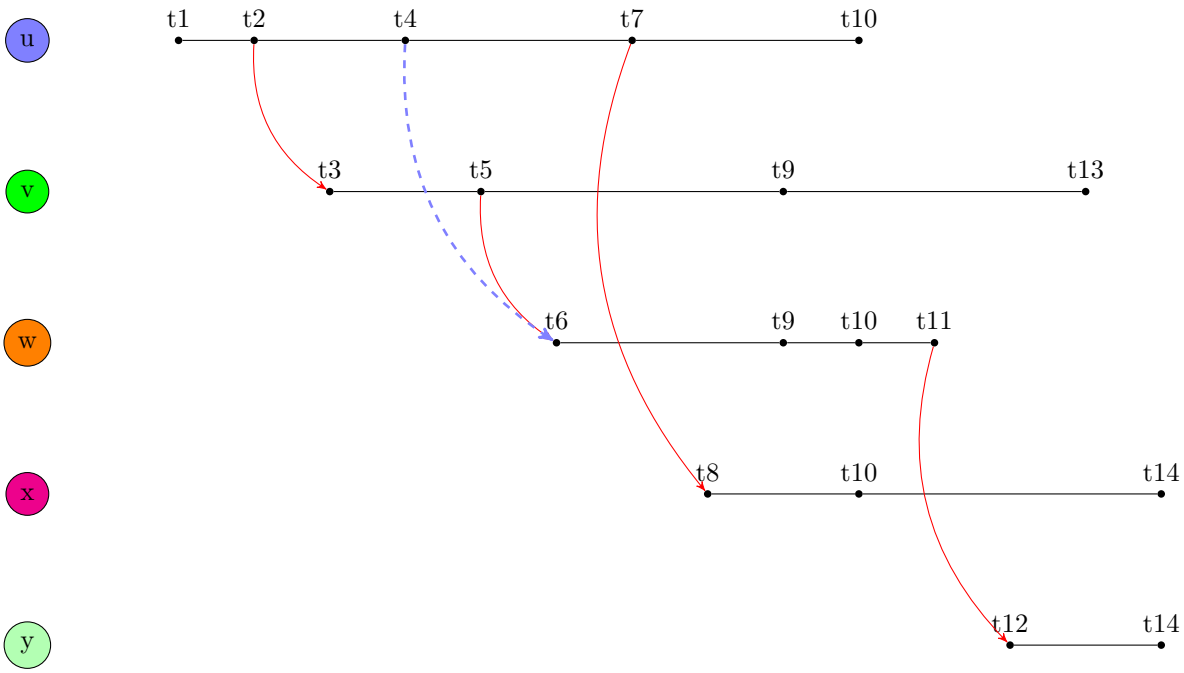


Figure 4: DAG of influence

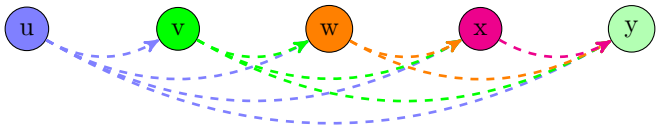
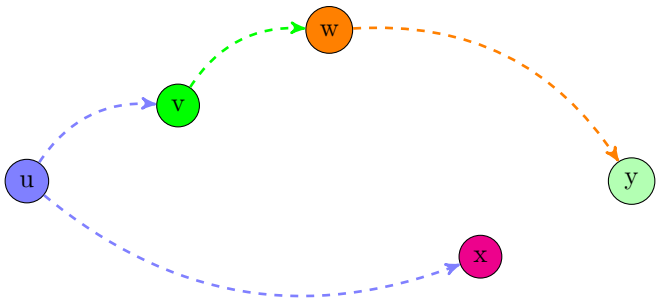


Figure 5: Influence tree



Notations	Description
$S = \{s_1, s_2, \ldots, s_n\}$	set of distinct sites
$G = \{g_1, g_2, \ldots, g_m\}$	set of well defined graphemes
S_{g_i}	set of sites inscribing a grapheme g_i
$T_{s_x}^{g_i} = \{t_0^{s_x,g_i}, \ldots, t_{T-1}^{s_x,g_i}\}$	set of distinct timestamps at which s_x inscribed g_i
$t_0^{s_x,g_i}$	earliest time of adoption of grapheme g_k by site s_x
$\Delta(t_j^{s_x,g_i}, t_k^{s_y,g_i})$	time difference between the $j-th$ and $k-th$ inscription of a grapheme g_i of sites s_x and s_y respectively
$\Delta_{\min}(t_i^{s_x,g_i}, t_0^{s_y,g_i})$	time difference between the earliest time of adoption of grapheme g_i by site s_y and the latest time of inscription by s_x before s_y, i.e. $\min(t_0^{s_y,g_i} - t_j^{s_x,g_i})$: $t_j^{s_x,g_i} \in \{t_0^{s_x,g_i}, \ldots, t_{T-1}^{s_x,g_i}\} \bigwedge t_i^{s_x,g_i} \leq t_0^{s_y,g_i}$
$dist[s_x, s_y]$	shortest straight line geographical distance between sites s_x and s_y
DAG_{g_i}	Directed Acyclic Graph of grapheme g_i
T_{g_i}	Inferred influence tree of g_i
$T = \bigcup_{g_i \in G} T_{g_i}$	Union graph (network) of the influence tress of all graphemes $g_i \in G$
$wt_{g_i}(s_x, s_y)$	strength of influence of site s_x on s_y for grapheme g_i
$wt(s_x, s_y) = \sum_{g_i \in G} wt_{g_i}(s_x, s_y)$	strength of influence of site s_x on s_y for all graphemes
$N(S, R)$	Network of sites S and their R dyadic relations

Table 1: List of notations

Acyclic Graph (DAG) with edges from every site preceding in time, to all other sites inscribing the same grapheme. Note, that this is the most basic assumption that can be made about any propagation of influence process, i.e. a site can only influence another site forward in time. For the DAG no weighting over the edges is assumed.

If a grapheme is inscribed by n distinct sites, then the influence DAG simply consists of $\binom{n}{2}$ edges directed from earlier inscribing sites to later inscribing sites. For example, the earliest inscribing can potentially influence all the rest of the $n-1$ sites inscribing the same grapheme, and so on. Also, this framework assumes that all timestamps are distinct. In case not, the parallel inscribing sites are not influencing each other.

Once the DAG is built, we can start extracting the edges of the influence tree from it based on a simple principle that the edges with higher weights are more likely to influence a site. The weighting of edges is inversely proportional to the difference in time and distance of inscriptions and sites respectively. This weighting scheme is explained in Section 4.2.2.

Influence propagation tree The tree is built by determining the most probable source of influence for each site. That is, considering all potential sources of influence (all sites that inscribed a grapheme before a given site inscribed it *for the first time*), pick the most probable (explained in Section 4.2.2) source for the given site for the given grapheme.

Starting from the site that inscribes a grapheme for the first time, determine the source of each site based on all the incoming edges into it. For site that inscribes for the first time there is no earlier known source site. So it is marked as the "innovator" and is the "root" of the tree. This, for example is u in Figure 5. For every other site the source is determined by the following procedure. For each site we consider all the incoming edges of influence into it from the DAG. The source site that has the highest weight (determined by the method described in Section 4.2.2) is chosen as the source of influence for the given site. This process is continued for every site inscribing a given grapheme. In the end we have the entire influence propagation tree of the grapheme.

Influence propagation graph Once the influence tree is built for each grapheme independently, those paths can be synthesized into a complex network of potential contacts through which the set of sites (communities) could have potentially interacted. This network is built by taking the union of all the influence trees built for each individual grapheme. One need to be aware of and making conscious choices about the following factors while taking this union of Maximum Likelihood trees.

- The directionality of the influence trees is ignored to create the union. Since in some influence tree one can anticipate s_x to s_y edges and in others the reverse.
- A directed graph can be built by respecting the directions of influence. Hence, in this case both (s_x, s_y) and (s_y, s_x) edges would exist incase of the scenario in the previous point.
- Weights of influence can be dealt with in several ways:
  - If a weighted union of trees is created, the weights of all the (s_x, s_y) edges from the trees are summed and same for all (s_y, s_x) edges.
  - Or, a unidirectional edge is created between s_x and s_y based on consensus of sum of weights in either direction. That is, if $\sum_{g_i \in G} wt_{g_i}(s_x, s_y) > \sum_{g_i \in G} wt_{g_i}(s_y, s_x)$, the resulting union graph has an edge (s_x, s_y) weighted by either $\sum_{g_i \in G} wt_{g_i}(s_x, s_y)$ or by $\sum_{g_i \in G} wt_{g_i}(s_x, s_y) - \sum_{g_i \in G} wt_{g_i}(s_y, s_x)$.
  - An undirected weighted network is created by by summing up all the (s_x, s_y) and (s_y, s_x) weights.

To summarize the above procedure, the influence propagation network is built through the following three steps.

DAG: For each grapheme, from each site with the earlier inscription time to all other sites with later inscription time, a potential contact of influence is constructed. The strength of each asymmetric influence tie is estimated on the basis of temporal and proximal closeness of inscriptions and sites respectively.

Tree: For each DAG, for each site the strongest source of influence is picked.

Network: All the trees are combined to construct a complex network of influence and interactions among all sites inscribing the set of graphemes.

Next we describe how the temporal and physical proximity are used to determine the strength of influence.

4.2.2 Strength of influence

The strength of influence is dependent on two simple factors: the temporal closeness of the grapheme inscription and the physical proximity of the pair of sites. Intuitively, the farther in time two sites inscribe a grapheme the less likely it is that they influenced each other. In the same vein sites that are physically far apart are less likely to influence each other in their writing scripts.

We consider the exponential distribution as the probability distribution that determines the strength of influence. The exponential waiting time distribution (or negative exponential distribution) is extensively used in literature for estimating diffusion models [5, 4, 6, 7]. It describes the time between events in a Poisson process, i.e. a process in which events occur continuously and independently at a constant average rate. Which is applicable in the case of monuments being enacted (and inscribed) at various times independent of each other. Informally, the influence of an earlier inscribing site on a later inscribing site is an inverse function of the difference in the timing of their inscription. Moreover, the influence decreases exponentially with increasing time difference.

In the next section we formally describe the propagation mechanism and give the algorithm for extracting the influence tree.

4.2.3 Exponential waiting time distribution for influence propagation

The influence path model is atomically build by determining asymmetric weight of influence using the exponential waiting time distribution. That is, the probability with which a site s_x inscribing at time $t_i^{s_x}$ a grapheme $g_i \in G$ can influence a site s_y inscribing, for the first time, at a later time $t_0^{s_y}$ is proportional to:

$$P_T^{g_i}(s_x, s_y) \propto \begin{cases} e^{-\beta_T \times \Delta(t_i^{s_x,g_i}, t_0^{s_y,g_i})} & if\ t_0(s_x,g_i) \leq t_i(s_x,g_i) < t_0(s_y,g_i),\ \beta_T > 0 \\ 0 & otherwise \end{cases}$$

where $\beta_T > 0$ is the *rate* or *survival* parameter for the influence to propagate.

Since, our model is not restricted to only one time of inscription, rather we take into consideration the entire timeline of inscriptions of a grapheme by a site, to determine when and with what strength one site could have influenced a site inscribing for the first time at a later time, we select the *latest* time a site inscribed the grapheme before another first time inscribing site to measure how influential the former is to the later site. Hence, the probability of influence is the function of the minimum time difference between of inscriptions between the two sites. That is,

$$P_T^{g_i}(s_x, s_y) \propto \begin{cases} e^{-\beta_T \times \Delta_{\min}(t_i^{s_x,g_i}, t_0^{s_y,g_i})} & if\ t_0(s_x,g_i) \leq t_i(s_x,g_i) < t_0(s_y,g_i),\ \beta_T > 0 \\ 0 & otherwise \end{cases}$$

where, β_T is the rate of propagation of influence. Then, the weight of influence between a pair of sites is:

$$P_T^{g_i}(s_x, s_y) = \begin{cases} \beta_T \times e^{-\beta_T \times \Delta_{\min}(t_i^{s_x,g_i}, t_0^{s_y,g_i})} & if\ t_0(s_x,g_i) \leq t_i(s_x,g_i) < t_0(s_y,g_i),\ \beta_T > 0 \\ 0 & otherwise \end{cases}$$

where, β_T can be estimated as;

$$\beta_T = \frac{1}{E[\Delta_{\min}(t_i^{s_x,g_i}, t_0^{s_y,g_i})]} \quad \forall s_x, s_y \in S^{g_i}$$

That is, β_T is the inverse of the expected value of the minimum time differences among all pairs of sites inscribing g_i.

Intuitively, β_T here is the *survival* parameter in the sense that if a random variable $X = \Delta_{\min}(t_i^{s_x,g_i}, t_0^{s_y,g_i})$ is the minimum time between inscriptions of a grapheme on a distinct pair of sites, then the expected value of all such outcomes estimates the overall rate at which propagation happens among the set of sites inscribing a grapheme over time.

The distance based propagation is modeled the same way as the time based one. That is,

$$P_D^{g_i}(s_x, s_y) = \begin{cases} \beta_D \times e^{-\beta_D \times dist[s_x,s_y]} & if\ t_0(s_x, g_i) \leq t_i(s_x, g_i) < t_0(s_y, g_i),\ \beta_D > 0 \\ 0 & otherwise \end{cases}$$

where,

$$\beta_D = \frac{1}{E[dist[s_x, s_y]]} \quad \forall s_x, s_y \in S$$

Notice, that though distance is a symmetric measure, influence is not. Here, in the distance model even though the influence is estimated based on proximity only. The directionality of influence still respects the time order, that is, one site can influence another only and only if the former inscribed the grapheme, for the first time, before the later site.

Assuming time and distance are independent explanatory variables for estimating the influence propagation, one can estimate the combined affect of these two variables on propagation by taking a product of them.

$$P_{T+D}^{g_i}(s_x, s_y) = P_T^{g_i}(s_x, s_y) \times P_D^{g_i}(s_x, s_y)$$

That is, if the earliest inscription date of a grapheme for one site is smaller than another site, then there is a potential that the former site influenced the later one in adopting the grapheme. The strength of that influence is proportional to product of the closest inscription times of the two sites and the physical distance between the two sites.

4.3 Model validation

One important question that needs to be addressed is how well the model emulates the reality. In this case, once we have inferred the potential networks of influence, how can we establish it aligns with actual relations among the sites? For this, we use the interpretations of the inscriptions. As explained in Section 3.2, parts of inscriptions, as interpreted by linguists narrate various forms of mutual relations among the communities [**?**]. The qualitative interpretation of each relation is beyond the scope of this analysis. Hence, for the relationship data we do not take into account the type or directionality of the relation.

We compare the inferred graph of influence to the relationship data to guage how well the model can approximate the influence among the sites. Since we build sparse trees for each grapheme, there are two issues that arise while comparing each tree to the ground truth: *a*) each grapheme is inscribed by only a small subset of sites, so each individual tree represents only a small sample of sites in the region that we have the relational data for; *b*) each influence graph is the minimal set of links among the sites inscribing the grapheme, hence each grapheme inference graph is very sparse as compared to the set of relations among sites. To counteract these problems, instead of comparing each individual graph of influence, we combine all the inferred graphs for each grapheme to build a graph that is more representative of the geographic region (representing a combined set of all sites inscribing the graphemes under consideration) and is not restricted to a single–source–of–influence type of tree graph. One byproduct of this procedure is that we end up with a symmetric graph instead of a directed one since for various graphemes individual influence directions could be contradictory.

5 Results

In this section we summarize the results of the models developed in this work. First we give some basic summary of the grapheme data. Then an overall outlook of the ritual data. We then report the results of this work on a subset of graphemes. Further on we compare the results to the relationship data to determine the accuracy of our models' estimates. Lastly, we guage the robustness of our models under the rate parameter perturbations.

5.1 Summary of basic statistics of the grapheme data

Summary of the basic counts of the grapheme data are given in table 2.

No. of inscriptions	73,359
No. of unique graphemes	956
No. of unique sites inscribing graphemes	?

Table 2: Basic counts from the grapheme data

5.2 Summary of the basic statistics of relationship data

Summary of the basic counts from the relationship data are given in Table 3. Table 4 gives the counts of

No. of distinct relationships	6
No. of distinct sites	79
No. of distinct sites with valid dates for relations	73
No. of all relationships	415
No. of relationships with valid dates	394
No. of all relationships with invalid dates	21
No. of distinct dyads in a relationship	143

Table 3: Basic counts from the relationship data

relationships of each type among the sites.

Relationships	No. of dyads	No. of distinct dyads
antagonistic	103	48
diplomatic	98	50
dynastic	22	15
kinship	27	18
nametag	37	25
subordination	25	21
unknown	82	63
Total	394	240

Table 4: Ground truth

Figure 6 depicts the relationship network among the 73 sites with valid date records. We do not distinguish between different relations in this analysis.

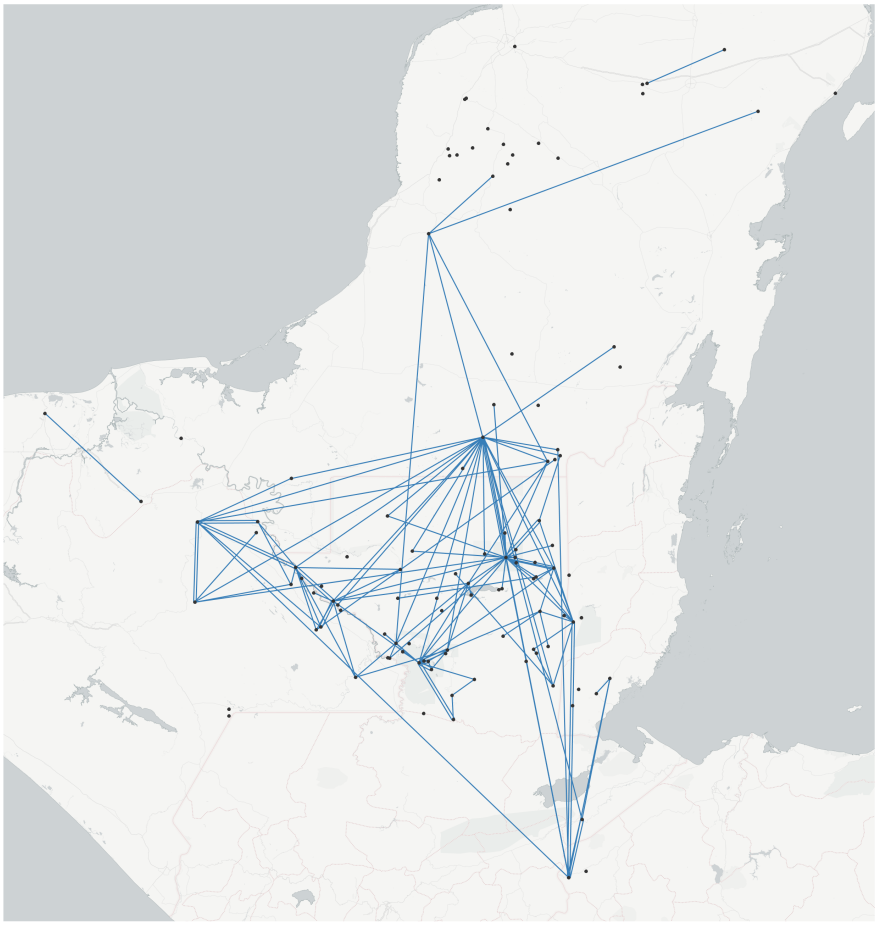


Figure 6: The relationship network

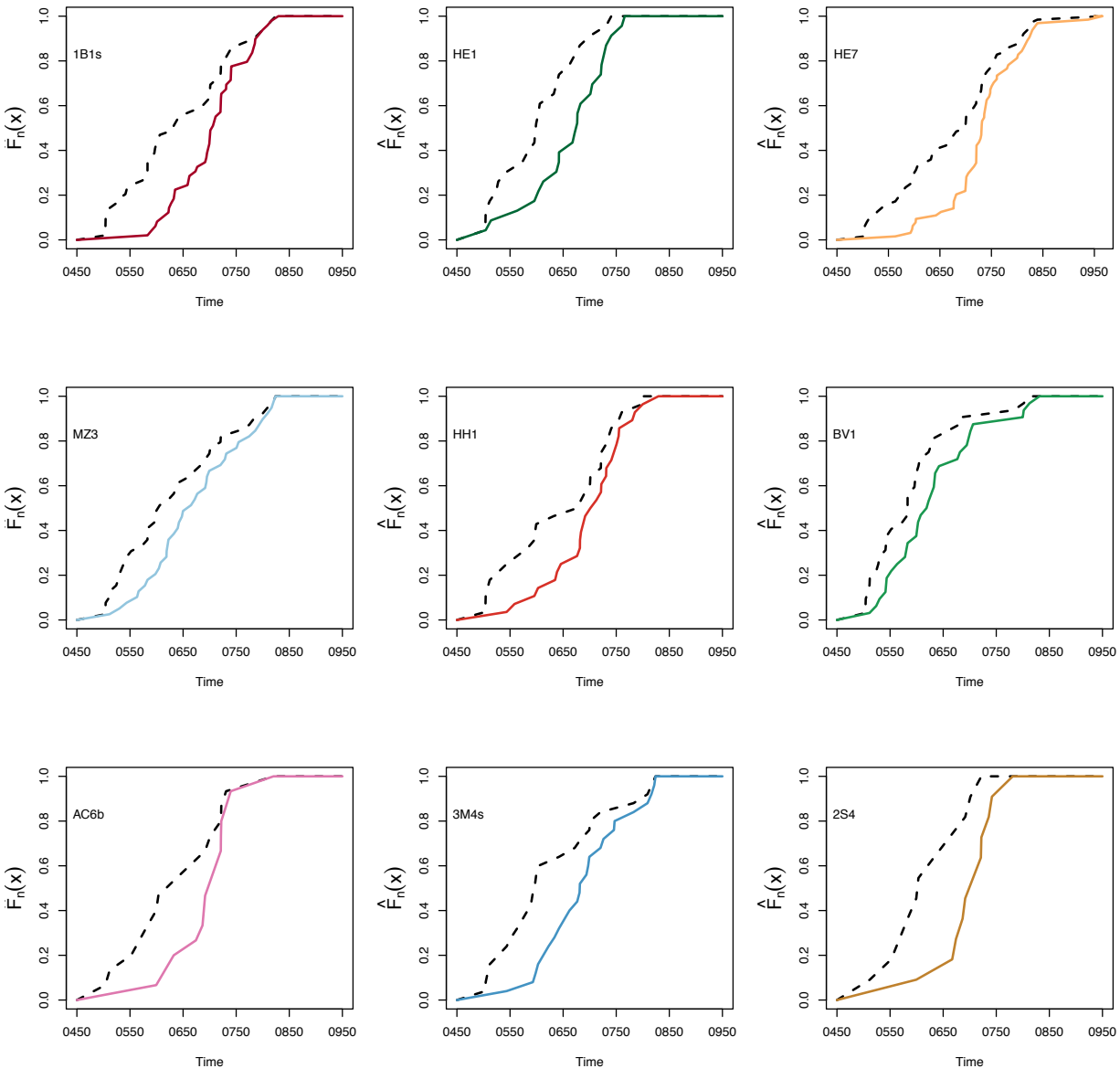
Sr.No.	Grapheme Code	Meaning	Usage	No. of unique inscriptions
1	HE7	he/she/it	syllabic	64
2	MZ3	?	syllabic	39
3	1B1s	?	syllabic	49
4	HH1	bone; captive	logographic	28
5	BV1	ajaw	logographic	32
6	AC6b	guard	logographic	15
7	3M4s	?	syllabic	25
8	2S4	?	syllabic	11
9	AL2b	agentive	?	26
10	AM8	?	logographic	13
11	XE3	?	syllabic	28
12	XQC	shield	logographic	23
13	ZX2	dawn; open	logographic	13
14	SM1	headband; paper, book	logographic	15
15	HE8	he/she/it	syllabic	15
16	ZG2s	?	syllabic	19
17	SCM	bone; captive	logographic	10
18	HE1	see, witness, observe	logographic	23

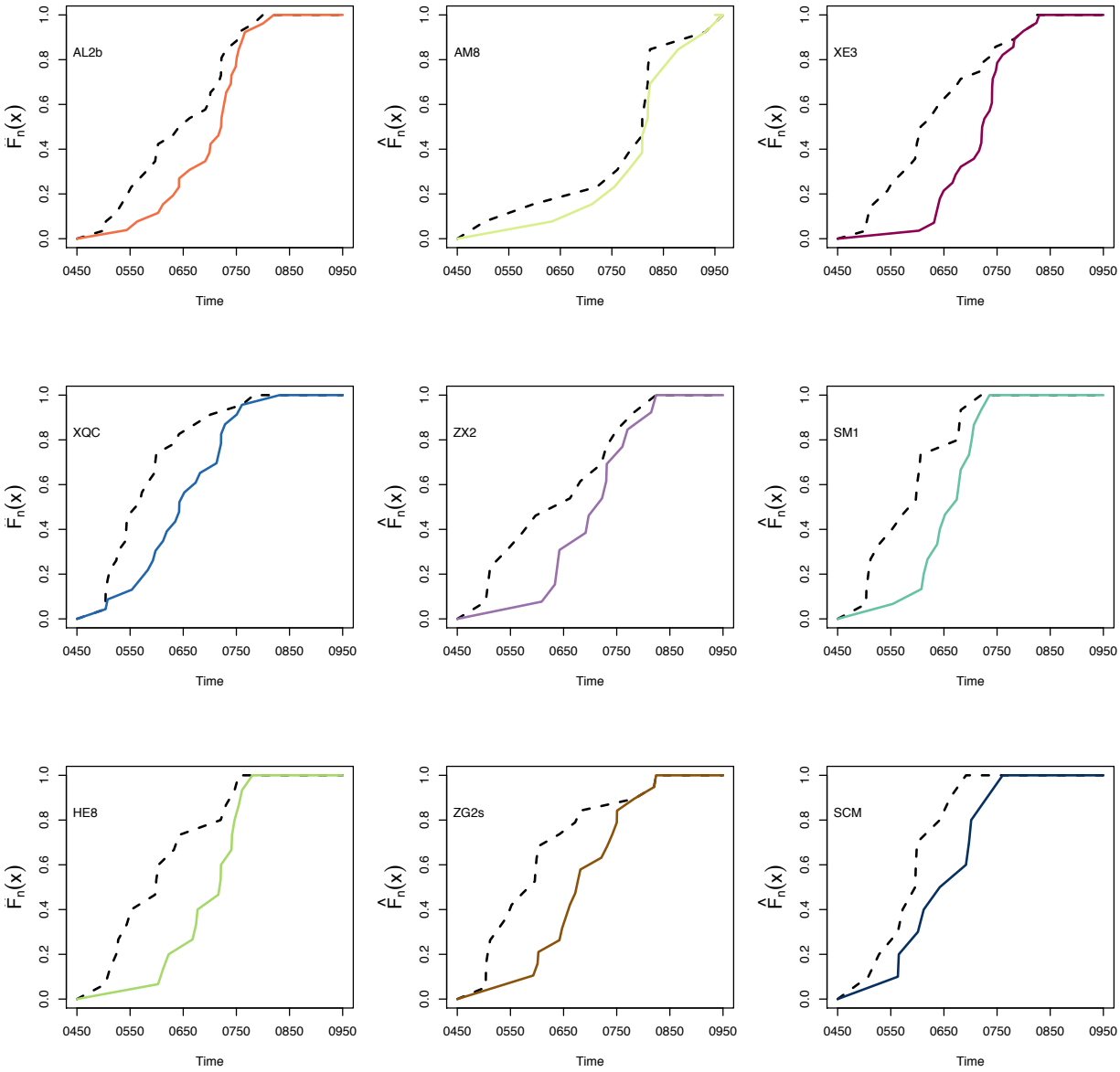
Table 5: Grapheme summary

5.3 Inferred Influence paths of graphemes

For brevity we show the results of the inference of influence paths procedure for a few selected graphemes. Following are inferred influence trees for 18 of the 956 graphemes. These graphemes were selected based on the significance of their contextual meaning by the domain experts. Table 5 gives the summary of the graphemes analyzed in this work. Figure 7 shows the s-curves of inscription over time for the 18 graphemes.

Figure 7: S–curves of graphemes

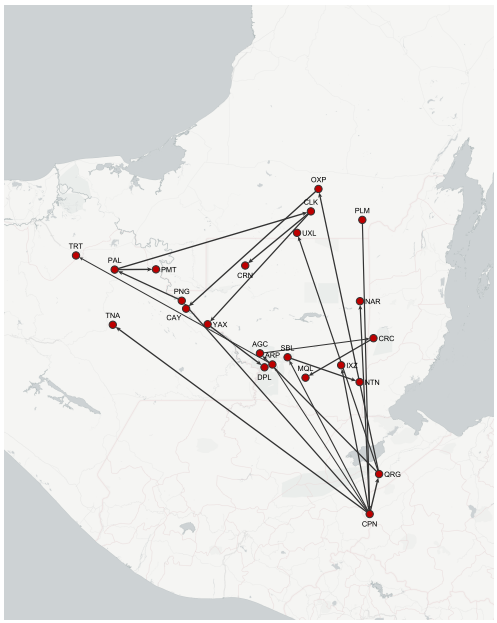




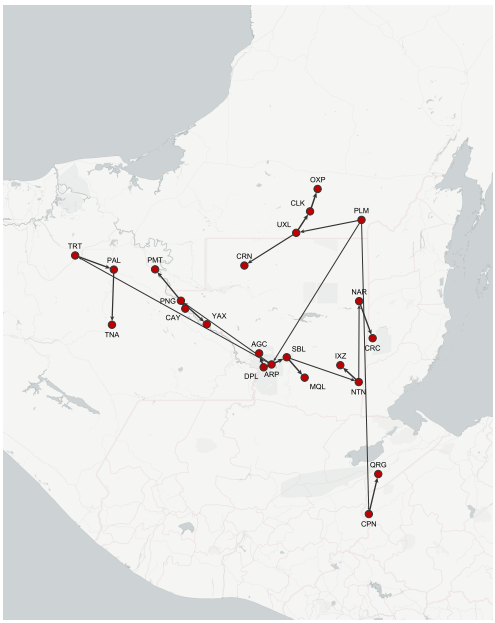
The following figures show the inferred paths of influence propagation for each of the 18 graphemes. For only one grapheme we show the resulting inferred trees for each model separately, that is, temporal model, distance model, and temporal and distance model combined. For all graphemes we show the temporal and distance modeled combined on the geographical map and the simple influence trees to get a sense of how influence propagation could have happened for each grapheme. In the figures, the direction of the edges indicate who influenced whom and thickness represent the relative weight of influence.

5.4 Model validation

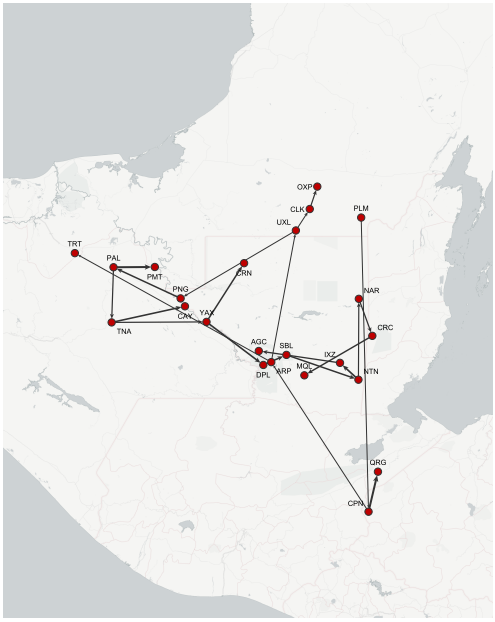
We compare the influence network of graphemes to the relationship data. Since the relationship data elucidates the actual interactions among the Maya communities, it is logical to compare the network inferred



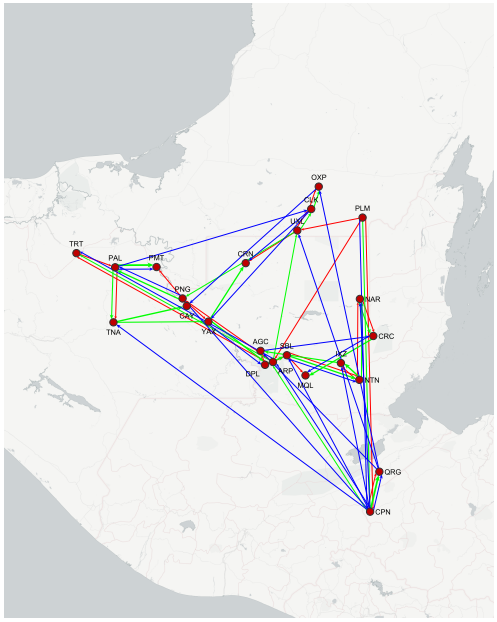
(a) Time



(b) Distance



(c) time and distance



(d) comparison time, distance, time+distance

Figure 9: HE1: the influence path graphs

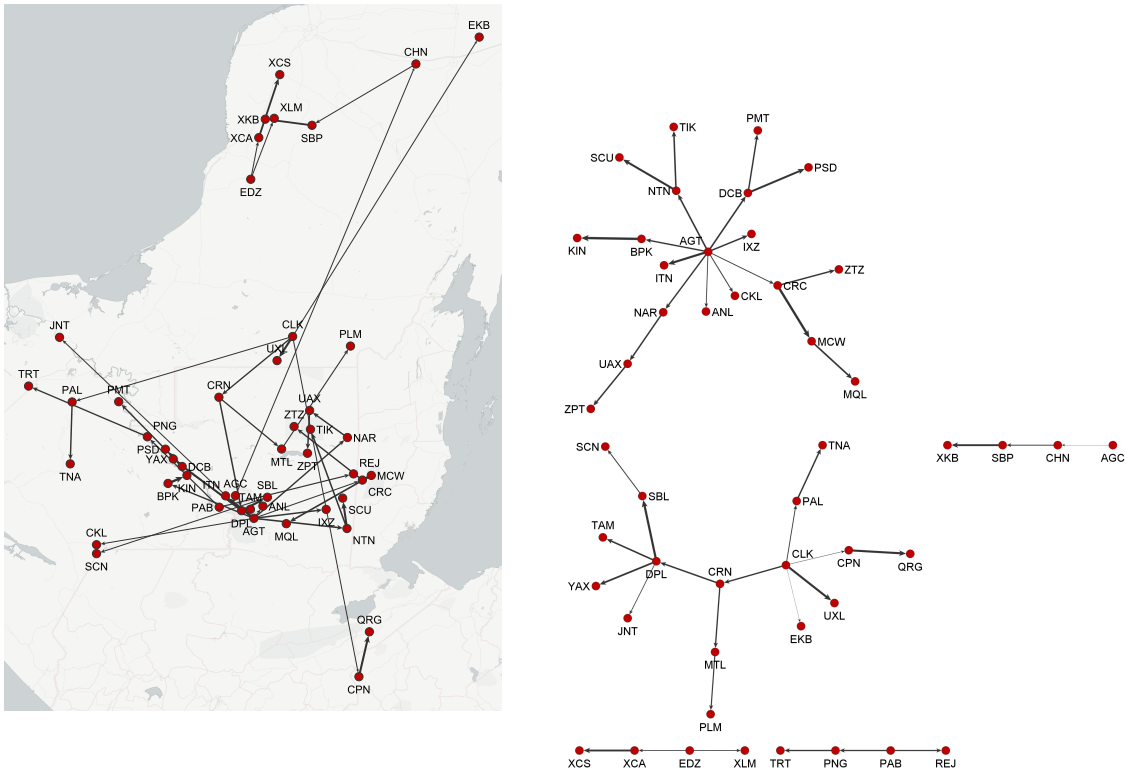


Figure 10: 1B1s: Influence paths - geo map and tree

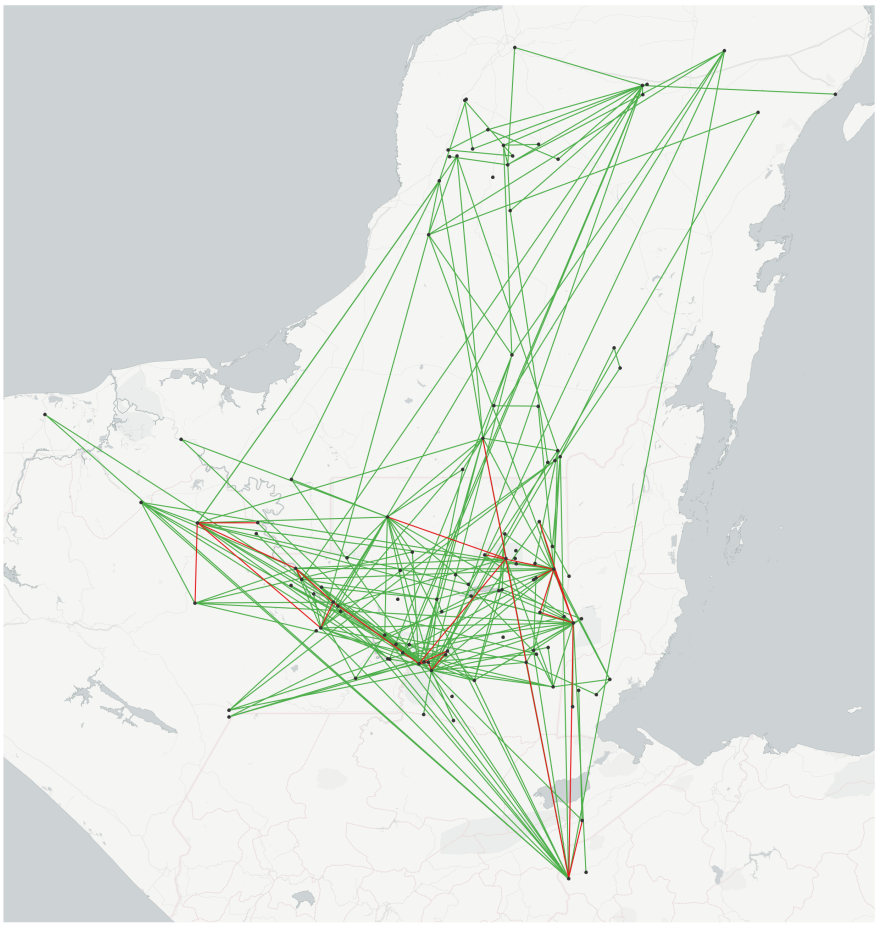


Figure 11: Complete inferred network with true edges in red

through this influence paths estimation to the relationship data. In Section 4.3 we state some of the limitations of this comparison. Figure 11 shows the combined network of influence edges for the 18 graphemes listed in Table 5.

First we combine the individual influence tree of each grapheme into one consolidated undirected graph. The union of trees is done as explained in Section 4.2.1. Figure 12 shows the relationship and influence edges on the geo map. Red edges: common in two networks, green edges: relationship network only,
blue edges: influence network only. Figure 11 shows the two graphs without the map.

Comparison of relationship network with grapheme network Here we evaluate how well our inferred graph fits the actual data. Table 6 summaries the comparison of the ground truth with the inferred network. The inference graph (union of the 18 graphemes stated above) represents the 102 sites inscribing those graphemes. The number of distinct inferred contacts (influence) among this set of sites is 253. Remember this graph is the combination of inference trees with the minimal number of edges. Hence, the overall density of this graph is very low. When compared to ground truth, the true relationships correctly identified my the inference model are 25 only. Hence, with this set of graphemes we do not

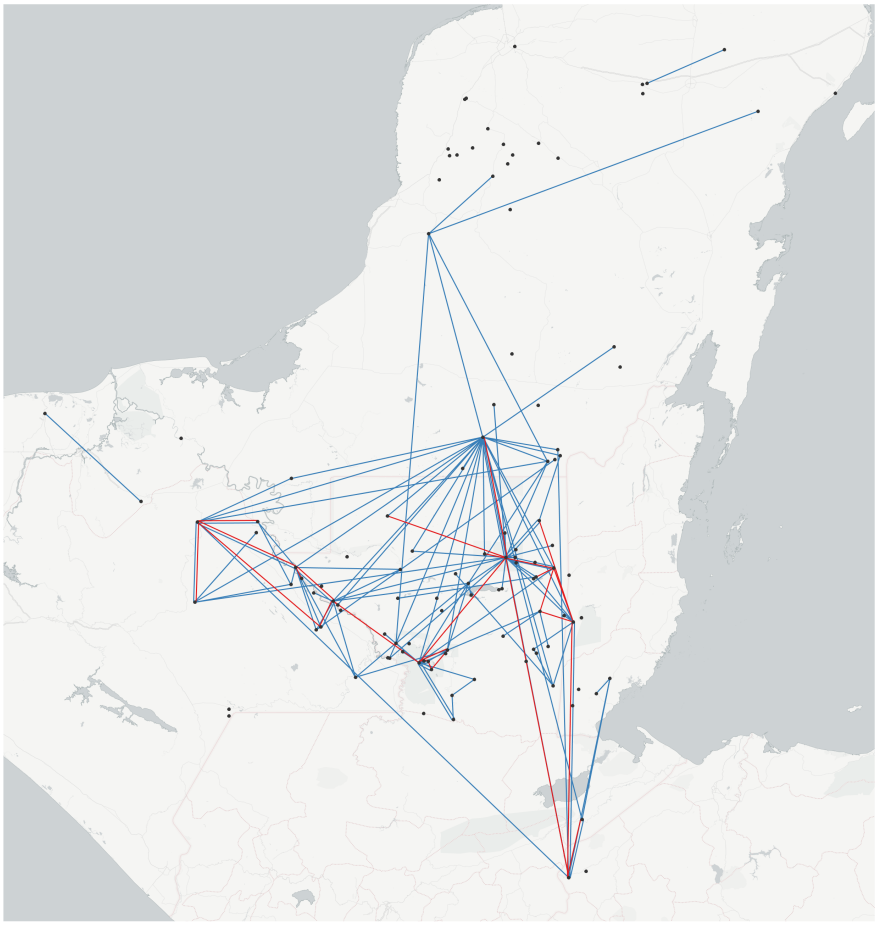


Figure 12: Relationship and correctly inferred edges on the geographical map

No. of sites in relationship network	73
No. of sites unique dyads relationship network	143
No. of sites in Inferred network	102
No. of unique dyads in inferred network	253
Density relationship network	0.0544
Density inferred network	0.003199
No. of common edges in ground truth and inferred network	25
Sites in relationship network also in inferred	62
No of edges in the induced network of matched sites	129
No of edges in the induced inferred network of matched sites	164
No of common edges in the induced network of matched sites and inferred network	25
Jaccard index of complete relationship and inferred network	0.067
Jaccard index of induced relationship and inferred network	0.0932
Ratio of matched to all inferred edges	0.0988
Ratio of matched to all inferred edges in the induced graph of matched sites	0.152

Table 6: Summary of relationship and influence network comparison

6 Discussion

6.1 Inferring complex influence propagation networks

Current results show that we achieve around 19% of accuracy when comparing the inferred network with the ground truth. There are two straightforward ways of improving accuracy of the current model.

6.1.1 Model based approach

Incorporating multiple sources of influence instead of limiting the source of influence to only one site. Our current model is based on the simplifying assumption that each site is influenced by only one site that has inscribed the grapheme before the site under consideration inscribes the same grapheme for the first time. The source site is picked based on the strength of influence defined in Section 4.2.2. Obviously this is a simplifying assumption. However, the challenge with selecting multiple sources of influence is that how to decide how many sources of influence can a site has? Potentially, any site inscribing a grapheme before the current site can potentially be a source of influence. A way to pick multiple sources is by some form of *thresholding*. For example, either a pre–fixed number of sources of influence can be set for each site. Or a cut–off on the strength of influence can be used to determine the likely sources of influence for each site. Both approaches require systematic setting of thresholds which may not always be straight forward.

Taking into account the limitations of the current model and the challenges of extending it, we propose the following enhancement to the model.

In the next iteration of the inference of influence modeling, each inscribing site can have more than one source of influence. This model is built iteratively in the form of $1 - DAG, 2 - DAG, \ldots, (n-1) - DAG$ for each grapheme.

Definition 1. $k - DAG$. A $k - DAG$, where $1 \leq k \leq n-1$, is a directed acyclic graph in which each node has atmost k sources of influence.

After each iteration, the unified graph of all grapheme DAGs is compared against the ground truth. Both the *precision* and *recall* of the inferred graph is measured. (Quick definitions: precision (also called positive predictive value) is the fraction of retrieved instances that are relevant, while recall (also known as sensitivity) is the fraction of relevant instances that are retrieved. Both precision and recall are therefore based on an understanding and measure of relevance.)

In the case of grapheme influence propagation:

$$precision = \frac{|<ground\ truth> \cap <inferred\ graph>|}{|<inferred\ graph>|}$$

$$recall = \frac{|<ground\ truth> \cap <inferred\ graph>|}{|<ground\ truth>|}$$

An alternative way of understanding precision and recall is:

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

where TP = True Positive, FP =False Positive, and FN =False Negative.

In simple terms, high precision means that an inference method returned substantially more relevant results than irrelevant ones, while high recall means that the method returned most of the relevant results. For example, a perfect precision score of 1.0 means that every result retrieved by the method was relevant (but says nothing about whether all the relevant edges were retrieved) whereas a perfect recall score of 1.0 means that all relevant edges were retrieved by the method (but says nothing about how many irrelevant edges were also retrieved).

Based on the precision and recall values we can decide whether to build the next (larger) DAG or to halt the process. Basically the decision is based on what our criteria of a "good" result is. As an example, one such criterion can be the F_1 score. In statistical analysis it is the measure of a method's accuracy. It considers both the precision and recall of the method. This score can be interpreted as a weighted average of the precision and recall, where an F_1 score reaches its best value at 1 and worst at 0. Mathematically it is represented as:

$$F_1 = 2 \times \frac{precision \times recall}{precision + recall}$$

This overall approach ensures that we retrieve best results with respect to the given ground truth while controlling for the edges that we do not have the any ground truth for. At the same time, this incremental approach aids us in circumventing the thresholding problem.

6.1.2 Data–driven approach

Another way of improving the accuracy of the inference method is by utilizing more graphemes. It is intuitive that the more data we have to build the graph, the better the accuracy of the inference model since more data will make the model richer. So far we have only worked with 18, most inscribed graphemes, out of the little over 900 in total. The question is if we add more graphemes then where to do we stop? The opposite extreme is to use all the graphemes. One small issue with this approach is that it is computationally expensive to infer diffusion trees for such a large set of graphemes. Although this process is fully parallelizable and hence not time inefficient. However, a bigger drawback of this approach is that with such a large input to the inference method we may get a result that is high in recall but low in precision, that is, all the true positive edges are retrieved but the number of false positives are proportionally very large. Or, we simply get something like a very dense clique like graph where almost all sites are mutually connected, making the results not very meaningful. Hence, building an influence graph with all the graphemes is not an interesting strategy. However, how many graphemes to sample and additionally how to sample the graphemes require some systematic approach.

One approach can be to build the graph by incorporating one grapheme at a time. This would result in a monotonically non–decreasing outcome in terms of true positive edges being retrieved[1]. That is, with

[1]A side note: recall will be monotonically non–decreasing but precision will not.

each additional graphemes the method performs better with respect to comparison to the ground truth but no worse.

Formally, this method can be described as follows. Let $G' \subseteq G$ be a non–empty subset of graphemes that have so far been considered for building the inference graph. Let $g_i \in G \wedge g_i \notin G'$ be a grapheme that has not yet been considered in building the inference graph. Let $M(N,T)$ be a function that extracts the matched edges between the ground truth network N and T - the union graph of influence trees of graphemes. Then,

$$M(N,T') = E'$$

where $T' = \bigcup_{g_i \in G'} T_{g_i}$ is the union graph of the inference trees of the graphemes from G' and E' is the set of edges that matched between the ground truth N and T'.

Proposition 1. $\forall g_i \in G \wedge g_i \notin G'$ the set of true edges recovered by applying the inference method on the set of graphemes $G' \cup g_i$ is at least as large as the set of true edges recovered by using only the set of graphemes G'.

Proof.

$$M(N,T') \leq M(N,T'')$$

The proof is straight forward. Will add later.

□

6.2 Robustness of model

In the first phase of this work, we have utilized 18 of the most frequently inscribed graphemes to infer the influence graph. Using those 18 graphemes, the inference method successfully recovered about 15% of the true relationship edges. In Section 6.1 we propose a couple of extensions to the model to improve the accuracy of our method. However, another aspect of the proposed approach that needs to be addressed is, how to sample the set of graphemes to apply to the inference procedure in a systematic way? And additionally, how big the sample size should be? The later question is addressed briefly in Section 6.1.2 with respect to the frequency based sampling. However, there are alternative approaches that can be tested to weigh the robustness of the proposed methods against various sampling strategies. Some of those approaches are listed below.

6.2.1 Sampling strategies

1. Random sampling: is the purest form of probability sampling. Each member of the population has an equal and known chance of being selected. When there are very large populations, it is often difficult or impossible to identify every member of the population, so the pool of available subjects becomes biased.

   Selecting a pre–defined number of graphemes uniformly at random from the entire set.

2. Systematic sampling: is often used instead of random sampling. It is also called an $N-th$ name selection technique. After the required sample size has been calculated, every $N-th$ record is selected from a list of population members. As long as the list does not contain any hidden order, this sampling method is as good as the random sampling method. Its only advantage over the random sampling technique is simplicity. Systematic sampling is frequently used to select a specified number of records from a computer file.

   Picking every $N-th$ grapheme or selecting the sample based on another pre–defined criterion. This strategy is closest to what we have employed in this work. That is, picking the most frequent set of graphemes in order.

3. Stratified sampling: is commonly used probability method that is superior to random sampling because it reduces sampling error. A stratum is a subset of the population that share at least one common characteristic. Examples of stratums might be males and females, or managers and non-managers. The researcher first identifies the relevant stratums and their actual representation in the population. Random sampling is then used to select a sufficient number of subjects from each stratum. "Sufficient" refers to a sample size large enough for us to be reasonably confident that the stratum represents the population. Stratified sampling is often used when one or more of the stratums in the population have a low incidence relative to the other stratums.

   Selecting graphemes based on the *syllabic* and *logographic* classification could be one such strategy. Selecting graphemes from a specific time window could be another approach.

4. Convenience/ Judgment sampling: is used in exploratory research where the researcher is interested in getting an inexpensive approximation of the truth. As the name implies, the sample is selected because they are convenient. This nonprobability method is often used during preliminary research efforts to get a gross estimate of the results, without incurring the cost or time required to select a random sample. In judgment sampling, the selection of the sample based on judgment. This is usually and extension of convenience sampling. For example, a researcher may decide to draw the entire sample from one "representative" city, even though the population includes all cities. When using this method, the researcher must be confident that the chosen sample is truly representative of the entire population.

   In the case of graphemes, considering only the largest connected component of the ground truth as the basis of grapheme selection. (Not sure how that would work though.)

6.3 Incorporating meta–data to the model

Incorporating supplemental information, such as, the frequency of inscriptions per site as an indicator of influence of the site.

6.4 Interpretation of the influence networks

Understanding the influence propagation process in terms of "linguistic attributes" of graphemes, such as, *logographic* and *syllabic* graphemes. So far we have focused on graphemes without taking into consideration their meaning and interpretations. It would be interesting to explore whether different types of graphemes, such as, logographic and syllabic exhibit differences in patterns of propagation of influence. Furthermore, certain meanings have been inscribed in more than one way, for example, the term "lord" can be inscribed using different graphemes. It would be interesting to explore if the trends of diffusion for various graphemes that means the same thing say something about influence among the sites.

6.4.1 Accounting for missing data in the ground truth

An interesting post session discussion point was the fairly common limitation of missing information in archaeological data. This study fits very well in such a paradigm. To mention just one anecdotal example from the Classic Maya data, site (?) has a high count of incoming edges in the ground truth but not many outgoing links due to the illegible and eroded monuments found at that site. It shows that although this site participated in many inter–community relations inscribed by other sites, there is no way to determine the counter relations it had with other sites. Hence, it is hard to gauge the social relations of this site. The by–product of this kind of missing information is that the ground truth is not complete and comparing with incomplete information the inference method may not be able to give us accurate evidence of interpersonal influence among the Maya communities. Therefore, one direction to explore in this study could be if we can infer some of those missing links. However, this analysis should be done in a future extension of the study as this question of inferring missing links could not be completely aligned with the direction of this work.

References

[1] Dana Angluin, James Aspnes, and Lev Reyzin. Inferring social networks from outbreaks. In *Algorithmic Learning Theory, 21st International Conference, ALT 2010, Canberra, Australia, October 6-8, 2010. Proceedings*, volume 6331 of *Lecture Notes in Computer Science*, pages 104–118. Springer, October 2010.

[2] Fred Brauer and Carlos Castillo-Chávez. *Basic Ideas of Mathematical Epidemiology*, pages 275–337. Springer New York, New York, NY, 2001.

[3] Fred Brauer, Pauline van den Driessche, and Jianhong Wu, editors. *Mathematical Epidemiology*, volume 1945 of *Lecture Notes in Mathematics*. Springer Berlin Heidelberg, 2008.

[4] Manuel Gomez-Rodriguez, D. Balduzzi, and B. Schölkopf. Uncovering the temporal dynamics of diffusion networks. In *Proceedings of the 28th International Conference on Machine Learning*, pages 561–568, Madison, WI, USA, July 2011. Omnipress.

[5] Manuel Gomez-Rodriguez, Jure Leskovec, and Andreas Krause. Inferring networks of diffusion and influence. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '10, pages 1019–1028, New York, NY, USA, 2010. ACM.

[6] Manuel Gomez-Rodriguez, Jure Leskovec, and Bernhard Schölkopf. Structure and dynamics of information pathways in online media. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, WSDM '13, pages 23–32, New York, NY, USA, 2013. ACM.

[7] Amit Goyal, Francesco Bonchi, and Laks V.S. Lakshmanan. Learning influence probabilities in social networks. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, WSDM '10, pages 241–250, New York, NY, USA, 2010. ACM.

[8] Adrien Guille, Hakim Hacid, Cecile Favre, and Djamel A. Zighed. Information diffusion in online social networks: A survey. *SIGMOD Rec.*, 42(2):17–28, July 2013.

[9] Matthew O. Jackson. *Social and Economic Networks*. Princeton University Press, Princeton, NJ, USA, 2008.

[10] David Kempe, Jon Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '03, pages 137–146, New York, NY, USA, 2003. ACM.

[11] Conrad Lee, Bobo Nick, Ulrik Brandes, and Pádraig Cunningham. Link prediction with social vector clocks. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13, pages 784–792, New York, NY, USA, 2013. ACM.

[12] Jure Leskovec, Andreas Krause, Carlos Guestrin, Christos Faloutsos, Jeanne VanBriesen, and Natalie Glance. Cost-effective outbreak detection in networks. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '07, pages 420–429, New York, NY, USA, 2007. ACM.

[13] Jessica Munson, Jonathan Scholnick, Matthew Looper, Yuriy Polyukhovych, and Martha J. Macri. Ritual diversity and divergence of classic maya dynastic traditions: A lexical perspective on within-group cultural variation. *Latin American Antiquity*, 27(1):74–95, 2016-03-03T00:00:00.

[14] Eldar Sadikov, Montserrat Medina, Jure Leskovec, and Hector Garcia-Molina. Correcting for missing data in information cascades. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, WSDM '11, pages 55–64, New York, NY, USA, 2011. ACM.