

# Probabilistically Inferring the Community Score

Bowen Hui (January 20, 2009)

## 1 The General Approach

There are several forms of *inference* which are generally classified as one of the following: *logical inference* and *probabilistic inference*. Logical inference uses traditionally logic and reasoning to draw conclusions. However, the only approach that is able to model uncertainty in a principled manner is probabilistic inference. Thus, this document focuses on probabilistic inference.

*Graphical models* is an intuitive and formal design approach that enables researchers to explicitly and concretely represent random variables and causal relationships in a domain. There are two components to graphical models: (i) a directed acyclic graph illustrating random variables as nodes and their causal relationships as links, and (ii) a numerical component that specifies the quantitative functions for each random variable. The graphical component corresponds to the *structure* of the problem, while the numerical component corresponds to the *parameters* of the model. Typically, the structure is handcrafted using domain expertise, although it can be learned empirically using extensive data. Parameters are sometimes handcrafted as well, but it is more common today to design data collection experiments and learn the parameters empirically. Therefore, it is crucial to develop a working model of the domain before conducting experiments, as the data collected from these experiments feed into learning the model parameters.

Probabilistic inference is carried out using fundamental principles of probability theory. The structure of a graphical model makes various simplifications possible (e.g., parts of the equation that are independent can “cancel out” of the equation). In addition, both approximate and exact algorithms to facilitate probabilistic inference have been heavily studied in the area of machine learning in Computer Science. Thus, developing working models for realistic domains is feasible.

An example of a general and expressive graphical model is a *dynamic Bayesian network*. This model is more powerful than its competitors because its inference algorithms exploit structural dependencies, which other models are unable to achieve. As such, this document illustrates how a community’s inferred score is modeled using a dynamic Bayesian network.

## 2 Steps in Model Construction

The general steps in building a model are outlined as follows:

1. repeat these steps until satisfied:
  - (a) handcraft the model structure and parameters using domain expertise
  - (b) simulation testing for semantic verification
2. real-user experiments for parameter acquisition for validation
3. re-do simulation testing with new parameters

### 2.1 The Role of Simulation Testing

The use of simulations for model checking is analogous to conducting a pilot study in empirical experiments. Simulations are conducted before (human) experiments take place, so they serve as a kind of pre-pilot study. The advantage that simulations have over pilot studies is that simulations do not require human participants.

The main reason for conducting simulations is to force researchers to explicitly state all assumptions and requirements of the model and follow-up studies. Often, a model description may make sense at a

high-level, but cannot be implemented because some steps are missing or the stated assumptions are not realistic. The deliverables from simulation experiments is an objective performance evaluation on a pre-determined set of metrics. Where appropriate, simulations serve as a comparative evaluation by running multiple methods and assessing their performance over the same metrics.

### 3 The Inference Task

The objective is to compute an “inferred score” of one community without directly surveying it. For the time being, let us concentrate on a group of similar communities, where the group consists of two communities denoted as  $A$  and  $B$ . In the simplest set-up, one could compute the inferred score of community  $A$ , given the surveyed score of community  $B$ , and potentially some other observations  $Obs1, Obs2$ , etc. A simple version of the interested task is expressed as follows:

$$Pr(ScoreA|ScoreB, Obs1, Obs2, \dots) \tag{1}$$

where  $ScoreA$  needs to be computed,  $ScoreB$  is given (i.e., observed), and  $Obs1, Obs2, \dots$  are additional available observations that influences  $ScoreA$ . Equation 1 is a naive version of the ultimate objective because it has no representation of time. For example, suppose at the beginning of the project, all the communities in one group are surveyed. Let us denote this point in time as time  $t = 0$ . With only two members in this community group means there are two values,  $ScoreA_0$  and  $ScoreB_0$  (the subscript denotes the time period). Two years later, community  $A$ 's score is surveyed, but  $B$ 's score is inferred. This provides the values  $ScoreA_1$  (i.e., surveyed score of community  $A$  at time  $t = 1$ ) and  $EstB_1$  (i.e., the *estimated* score of community  $B$  at time  $t = 1$ ). Reversing the sampled community two years later, community  $B$ 's score is surveyed, but  $A$ 's score is inferred. At this time  $t = 2$ , the inference task becomes:

$$Pr(ScoreA_2|ScoreB_2, Obs1_2, Obs2_2, \dots, ScoreA_1, EstB_1, ScoreA_0, ScoreB_0) \tag{2}$$

where all the variables behind the “|” bar is observed. The corresponding graphical model describing Equation 2 is illustrated in Figure 1.

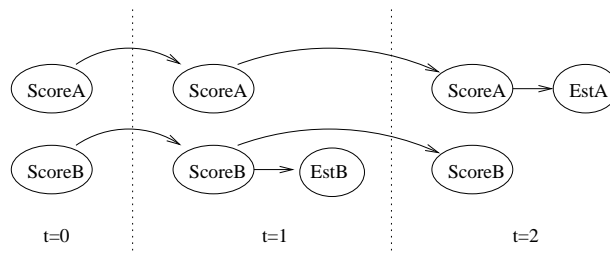


Figure 1: The graphical representation of Equation 2 without illustrating the variables  $Obs1, Obs2$ , etc. Note that the true score (surveyed or unobserved) from the previous time period persists over time so that it influences the corresponding score at the next time period.

In general, the task is to compute the inferred score of one community at the current time, given the surveyed score of another and other possible observations at the current time, plus the *history* of all the scores from the past. In general, the inference formula at time  $t$  is:

$$Pr(ScoreA_t|ScoreB_t, Obs1_t, Obs2_t, \dots, ScoreA_{t-1}, EstB_{t-1}, EstA_{t-2}, ScoreB_{t-2}, \dots, ScoreA_0, ScoreB_0) \tag{3}$$

where  $ScoreA_t$  needs to be inferred and  $ScoreB_t$  is surveyed,  $Obs1_t, Obs2_t, etc.$  are observations,  $ScoreA_{t-1}$  was surveyed and  $EstB_{t-1}$  was inferred at the previous time period,  $EstA_{t-2}$  was inferred and  $ScoreB_{t-2}$  was surveyed two time periods ago, the pattern continues to alternate until the initial scores  $ScoreA_0$  and  $ScoreB_0$  were surveyed at the beginning of the project.

## 4 Modeling Community Relationships

In addition to the model in Figure 1, other causal relationships may exist between communities that influence their scores. Consider the following examples:

- the scores of  $A$  and  $B$  are “correlated” (positively or negatively), such that a random variable influencing  $ScoreA$  will also influence  $ScoreB$
- $A$  is “ambitious”, such that, knowing its own score from the previous survey (estimated or sampled) will encourage  $A$  to try to do better
- $A$  and  $B$  are “competitive”, such that, if  $A$  knows  $B$  is better, then  $A$  will try to do better (and vice versa)

These relations are by no means exhaustive; domain expertise is needed in order to define all the causal relationships that are relevant to determining a community’s score over time.

The three example relationships suggest (perhaps naively) that the dynamics across communities can be captured using information only from the previous time period and the current time period. In other words, older history about the domain is not relevant, i.e., the computations in the current time period is *independent* of older history given the information only from the previous time period. If this is true, Equation 3 can be further simplified. Following this assumption, the graphical models in Figures 2-4 illustrate each of these relationships respectively. Each figure is described in the respective caption.

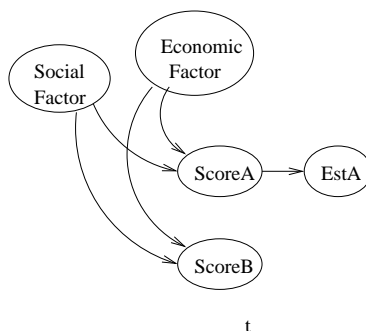


Figure 2: The scores of two communities are “correlated” via some economic or social factor. These factors influence the true score  $ScoreA_t$  and the survey score  $ScoreB_t$ . Variables and relationships from the previous time periods are omitted for clarity.

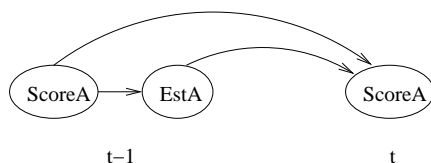


Figure 3: Community  $A$  is “ambitious”, where knowing the inferred score  $EstA_{t-1}$  from the previous time period influences  $A$ ’s current score,  $ScoreA_t$ . This relationship describes the dynamics over two time periods,  $t - 1$  and  $t$ . Variables and relationships from the previous time periods are omitted for clarity.

## 5 The Quantitative Component: An Example

As mentioned above, the quantitative component in a graphical model requires that each node specifies a distribution given its parent nodes. Consider the nodes in Figure 3 as an example. These variables are  $ScoreA_{t-1}$ ,  $EstA_{t-1}$ , and  $ScoreA_t$ . In particular,  $ScoreA_{t-1}$  and  $EstA_{t-1}$  are parent nodes of  $ScoreA_t$

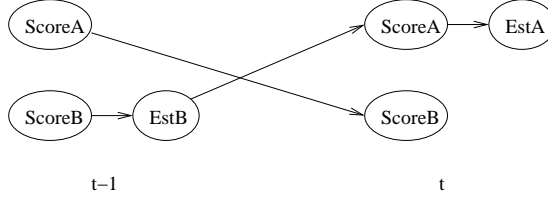


Figure 4: Communities  $A$  and  $B$  are “competitive”, where knowing  $A$ ’s previously surveyed score  $ScoreA_{t-1}$  influences  $B$ ’s current score,  $ScoreB_t$ , and knowing  $B$ ’s previously inferred score  $EstB_{t-1}$  influences  $A$ ’s current score (and hence, the current score to be inferred,  $EstA_t$ ). This relationship describes the dynamics over two time periods,  $t - 1$  and  $t$ . Variables and relationships from the previous time periods are omitted for clarity.

because they have links pointing to  $ScoreA_t$ . Since the values of these variables are defined as letter grades, these are *discrete* variables each with 5 distinct values. Thus, the quantitative components of these variables are represented as tables.<sup>1</sup> In particular, the quantitative component associated with  $EstA_{t-1}$  defines  $Pr(EstA_{t-1}|ScoreA_{t-1})$ , which can be defined by the *conditional probability table* in Table 1.

$ScoreA_{t-1}$	$EstA_{t-1} =$	A	B	C	D	E
A		$p_1$	$p_2$	$p_3$	$p_4$	$p_5$
B		$p_6$	$p_7$	$p_8$	$p_9$	$p_{10}$
C		$p_{11}$	$p_{12}$	$p_{13}$	$p_{14}$	$p_{15}$
D		$p_{16}$	$p_{17}$	$p_{18}$	$p_{19}$	$p_{20}$
E		$p_{21}$	$p_{22}$	$p_{23}$	$p_{24}$	$p_{25}$

Table 1: A table of probabilities (abstractly shown as  $p_1, \dots, p_{25}$ ) describing the probability that  $EstA_{t-1}$  takes on a certain letter grade, given each possible parent value defined by  $ScoreA_{t-1}$ . For example,  $p_5$  represents  $Pr(EstA_{t-1} = E|ScoreA_{t-1} = A)$ .

Another relationship illustrated in Figure 3 is  $Pr(ScoreA_t|ScoreA_{t-1}, EstA_{t-1})$ , which is the quantitative component associated with  $ScoreA_t$ . A table similar to Table 1 can be defined in this case. However, rather than a 5x5 table, this relationship requires a 25x5 table, because there are 25 combinations of possible pairs of values for the two parent nodes  $ScoreA_{t-1}$  and  $EstA_{t-1}$  in this case. It is easy to see from this example that larger models become difficult to scale.

<sup>1</sup>In contrast, the quantitative component of a continuous variable is a functional description of a distribution.