Is Guilt by Association a Bad Thing?

Aram Galstyan and Paul R. Cohen
USC Information Sciences Institute
4676 Admiralty Way, Suite 1001
Marina del Rey, CA, 90292, USA
{galstyan,cohen}@isi.edu

Keywords: Suspicion Scoring, Guilt by Association.

Abstract

In this paper we study a classification model that mimics guilt by association. We consider a population consisting of known adversaries, covert adversaries, and benign individuals. In our model we associate a suspicion score with each individual. The scores for the covert and benign populations are initially set to 0, while the known adversaries have a fixed score of 1. The scores change dynamically as individuals meet each other. An individual will be classified as an adversary if his score exceeds a certain threshold. We demonstrate analytically and empirically that a method for assigning suspicion scores, and hence the classification procedure itself, can be very sensitive to its parameters. This suggests that guilt-by-association models, and classification procedures based on them, should be treated very carefully and analyzed for robustness to their parameters.

1. Introduction

“Guilt by association” is widely understood concept in classification problems. It is used in one form or another for community finding in social networks (Clauset 2004, Newman and Girvan 2004), studying covert networks (Dombroski 2003; Krebs 2001), and relational classification (Macskassy and Provost 2003; Neville and Jensen 2000). In this paper, by “guilt by association” we mean a class of models in which an individual *i* is assigned a score $s_i(t)$ that depends on the scores of the individuals *j*, *k*,…,*m*, with whom *i* associates. Scores change over time as new associations between individuals are formed, so we are interested in models like this:

$$s_i(t+1) = \phi(s_j(t), s_k(t), ..) \qquad (1)$$

If ϕ is monotonic, then over time all scores in a group of associated individuals will increase without bound. As each individual’s score increases, it increases the scores of associated individuals at the next time step, and so on, in a mutually reinforcing way. Here we consider the class of guilt-by-association models in which scores are bounded by the [0,1] interval, and for whom the steady-state average score is less than or equal to 1. The following model has these properties:

$$\phi = s_i(t) - \beta s_i(t) + \alpha(1 - s_i(t)) z(s_j(t), s_k(t), ..) \qquad (2)$$

In this model *z* is a function of the scores of the individuals with whom *i* associates. The parameters α and β represent how much effect *z* has on *i*’s score, and a rate of decay of *i*’s score, respectively. The main result of this paper is that one’s ability to classify individuals on the basis of their scores is very sensitive α and β.

To make this result concrete, imagine one is trying to classify individuals as terrorists or non-terrorists. One obtains a time series of transactions: On Monday, a meeting took place between individuals 17, 23, and 48; on Tuesday, individual 48 met with individuals 17 and 91, and so on. One increases the suspicion scores of the individuals at meetings, proportional to some function of all the scores at the meeting, scaled by α. One decreases the suspicion score of every individual, whether or not they attend a meeting, proportional to β. After a number of iterations and meetings one ranks the individuals by their suspicion scores and guesses that the top-ranked individuals are terrorists. It turns out that the accuracy of this procedure is very sensitive to α and β.

One might argue that the model in Equation (2) is just a bad model, and that there are better ways to assign suspicion scores that are not so sensitive to their parameters. Our point is not that the model is a good one, but that we should analyze guilt-by-association models, and classification procedures based on them, for robustness to their parameters. That said, we think the model in Equation (2) has elements that will be found in all guilt-by-association models: There must be a term that represents the suspiciousness of associates (z in Equation (2)) and the effect of this term will always be scaled by some function (in Equation (2) it is a linear scaling by α). Not all models will be bounded by the interval [0,1], but there are strong reasons to want to have this requirement: If scores increase without bound then an individual's score must somehow be standardized to account for how long the individual has been "alive," otherwise a high score might signify nothing more than seniority. Similarly, classification based on suspicion scores is possible only if some individuals can have higher scores than others, so the steady-state average score cannot be the maximum score, and some decay parameter such as β in Equation (2) is required.

In the following sections we present our guilt-by-association model in more detail and then report analytical and empirical results that demonstrate how classification accuracy is very sensitive to the model parameters.

2. Model

Let us consider a population consisting of three types of individuals: There are N_a known adversaries, N_c covert adversaries, and N_b benign individuals. Assume that we can meetings between these individuals. For simplicity, we assume that in each meeting there are only two participants. Meetings between all kinds of individuals are observed. However, on average, covert adversaries are more likely to meet with known adversaries than with benign individuals. Our task is to identify the covert adversaries using this record of contacts.

We suggest the following guilt by association scheme for classifying individuals: All the known adversaries are assigned a permanent suspicion score s=1, all other individuals are assigned an initial score s=0 that will change as they meet with other individuals. Specifically, if an individual with score s_1 meets with an individual with a score s_2, then his score will be updated as follows:

$$s_1 \to s_1 + \alpha(1 - s_1)s_2 \tag{3}$$

In addition to the incremental growth in the suspicion score, we will also have it decay at rate β. Thus, the dynamics of suspicion scores is described by:

$$s_i(t+1) = s_i(t) - \beta s_i(t) + \alpha(1 - s_i(t))z \tag{4}$$

Here z is the score of the individual that the i-th individual has met at time t: if the individual does not meet anyone at time t then z=0. Generally speaking, z is a random variable with a certain distribution that we will specify in the next section. It is easy to see that this distribution will be different for benign individuals and covert adversaries since the rates at which they meet with various types of individuals are different. The main question is whether this difference in the meeting pattern will produce distributions of suspicion scores that can be used to classify covert adversaries as such.

If an individual has a suspicion score greater than a threshold, we will classify that individual as a covert adversary. In the experiments that follow, the threshold is always set to the value that gives the highest classification accuracy. In this way we can study the effects of α and β on classification accuracy without worrying about effects due to the threshold.

3. Statistical Analysis

The success of a classification algorithm that uses suspicion scoring depends crucially on how these scores evolve in subpopulations. As we already mentioned, the difference in the evolution of the scores in subpopulations is due to the difference in the meeting patterns between various types of individuals. To be specific, let us consider the case where meetings between individuals are governed by a random Poisson process. Specifically, we assume that each covert individual meets with known adversaries, covert adversaries, and benign individuals with rates C_a, C_c and C_b, respectively. Similarly, the benign individuals meet with known adversaries, covert adversaries, and benign individuals with rates B_a, B_c and B_b.

Further, let $P_c(s,t)$ be the probability distribution that a randomly chosen covert individual at time t has a score s. We define $P_b(s,t)$ similarly for the benign populations (note that $P_c(s,t)$ and $P_b(s,t)$ are simply the average fraction of covert and benign individuals with score s at time t). Then the variable z in Equation 4 can be easily shown to have the following distribution for each subpopulation:

$$p^c(z) = (1 - C_a - C_b - C_c)\delta_{z,0} + C_a\delta_{z,1} + C_bP_b(z,t) + C_cP_c(z,t) \tag{5}$$

$$p^b(z) = (1 - B_a - B_b - B_c)\delta_{z,0} + B_a\delta_{z,1} + B_bP_b(z,t) + B_cP_c(z,t) \tag{6}$$

Given the distribution for z, we can derive the stochastic Master equation for the evolution of the score distributions using the continuous time limit of Equation 4:

$$\begin{aligned}\frac{\partial P_c(s)}{\partial t} &= (\beta - C_a - C_b - C_c)P_c(s) + \beta s\frac{\partial P_c(s)}{\partial s} + \frac{C_a}{1-\alpha}P_c\left(\frac{s-\alpha}{1-\alpha}\right) + \\ &+ \int ds'\frac{C_c}{1-\alpha s'}P_c(s')P_c\left(\frac{s-\alpha s'}{1-\alpha s'}\right) + \int ds'\frac{C_b}{1-\alpha s'}P_b(s')P_c\left(\frac{s-\alpha s'}{1-\alpha s'}\right),\end{aligned} \tag{7}$$

$$\begin{aligned}\frac{\partial P_b(s)}{\partial t} &= (\beta - B_a - B_b - B_c)P_b(s) + \beta s\frac{\partial P_b(s)}{\partial s} + \frac{B_a}{1-\alpha}P_b\left(\frac{s-\alpha}{1-\alpha}\right) + \\ &+ \int ds'\frac{B_b}{1-\alpha s'}P_b(s')P_b\left(\frac{s-\alpha s'}{1-\alpha s'}\right) + \int ds'\frac{B_c}{1-\alpha s'}P_c(s')P_b\left(\frac{s-\alpha s'}{1-\alpha s'}\right)\end{aligned} \tag{8}$$

The initial conditions are $P_c(s,0)=\delta_{s,0}$ and $P_c(s,0)=\delta_{s,0}$. Note that the accuracy of classification can be expressed through the overlap of these distribution functions. Namely, if we choose a threshold S_{thr} for classification, then the fraction of False Positives (FP) and False Negatives are

$$FP = \int_{S_{thr}}^{1} ds P_b(s)$$

$$FN = \int_{0}^{S_{thr}} ds P_c(s)$$

While the master equations (7,8) can be solved readily through numerical integration, their analytical solutions are not available in the most general case. Instead, in this paper we prefer to work its first moment that describes the evolution of the average scores in both subpopulation. The rational behind this is that the difference in the average scores between two populations will be indicative of the measure of the overlap of corresponding distribution.

We now examine how the average suspicion score over two sub-populations changes with time. Let $S_c(t)$ and $S_b(t)$ be the suspicion score averaged over covert and benign populations at time t. The equations for $S_c(t)$ and $S_b(t)$ are obtained by multiplying the corresponding master equations by s and integrating over s. The resulting equations read:

$$\frac{dS_c}{dt} = -\beta S_c + \alpha(1 - S_c)(C_a + C_c S_c + C_b S_b) \qquad (9)$$

$$\frac{dS_b}{dt} = -\beta S_b + \alpha(1 - S_b)(B_a + B_c S_c + B_b S_b) \qquad (10)$$

Equations (9,10) are a coupled nonlinear system with the coupling coefficients given by the cross-population meeting rates C_b and B_c. In the case $C_b = B_c = 0$ the evolution in each sub-system is independent. Note also, that these equations could be obtained directly from Equation 4 by averaging over z using the distribution functions for z, Equations 5-6.

To test our analytical model, we carried out simulations of suspicion scoring scenario with two subpopulation of 1000 benign and 1000 covert individuals. The rates used were $C_c = 0.002$, $B_b = 0.002$, $C_a = 0.001$, $B_a = 0.0002$, $C_b = B_c = 0.0002$. In Figure 1 we plot the time evolution of the suspicion scores averaged over sub-populations for simulation and the numerical solutions of equations 9 and 10 for two different values of the parameters α and β. We did not average the simulation results over many runs in order to account for fluctuations. One can see that the analytical solution agrees very well with the simulation results. As it can be seen from Fig. 1, after some transient time (that varies with the choice of the parameters) the average scores evolve into a steady state. What is important is that *the difference between the steady state values for covert and benign populations depends on the choice of the model parameters* α *and* β.

To study this dependence, we note that these steady state values are the solution of following system of non-linear equations:

$$-\beta S_c + \alpha(1 - S_c)(C_a + C_c S_c + C_b S_b) = 0 \qquad (11)$$

$$-\beta S_b + \alpha(1 - S_b)(B_a + B_c S_c + B_b S_b) = 0 \qquad (12)$$

Note that the steady state values depend on the ratio β/α only. However, in the following we will study the dependence on these parameters individually.

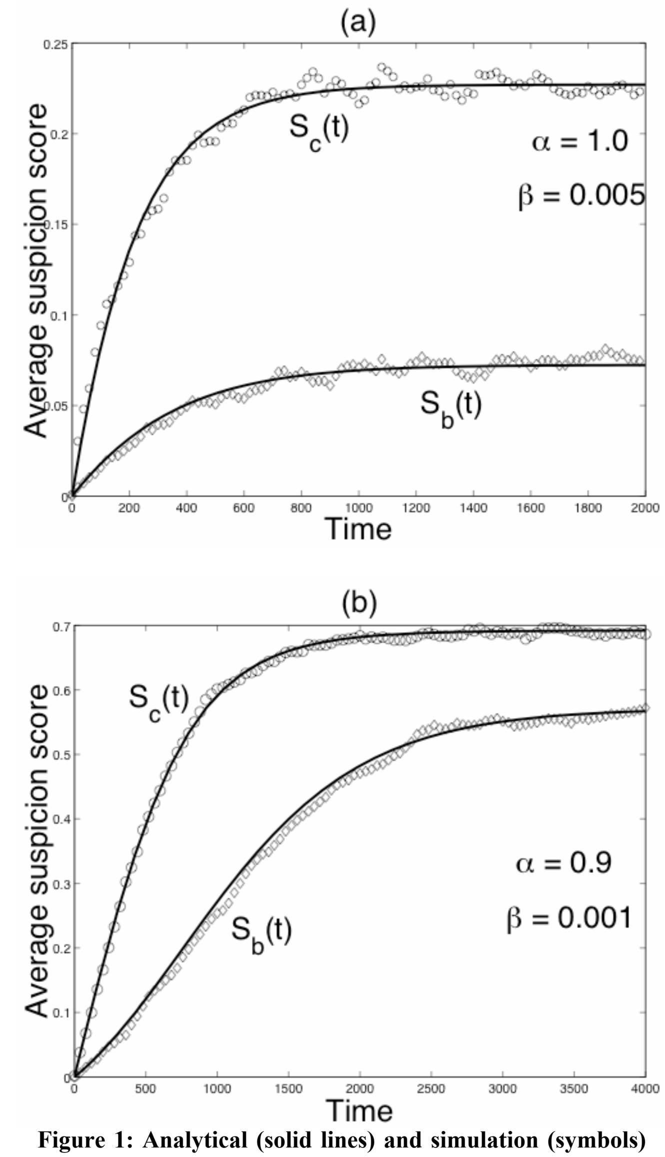


Figure 1: Analytical (solid lines) and simulation (symbols) results for the evolution of average suspicion score for each population.

Again, in the most general case, the solutions of the steady state equations have to be obtained numerically.

We use these solutions to examine how the difference between average scores in two subpopulations $\Delta S = S_c - S_b$ behave for various choices of parameters α and β.

In Figure 2 (top) we plot this difference versus decay parameter β for three different values of α. Remarkably, the dependence is highly non-monotonic with a single maximum at a certain value of β that shifts to right as α increases. This can be understood as follows: For a zero decay $\beta = 0$, the average score in both sub-populations will reach $s = 1$. The difference will be only in the transient time required to reach this maximum (e.g., longer for the benign population). Hence, the score difference will be zero. In the other extreme when the decay rate is very high, the scores of individuals decay exponentially fast with a rate $1/\beta$, so that the average scores in both populations will be close to zero again. Hence, ΔS should reach at least one extremum. Remarkably, the maximum seems to be the most pronounced (e.g., high and narrow) for the smaller values of α. This can also be seen in the surface plot of Fig. 2(bottom).

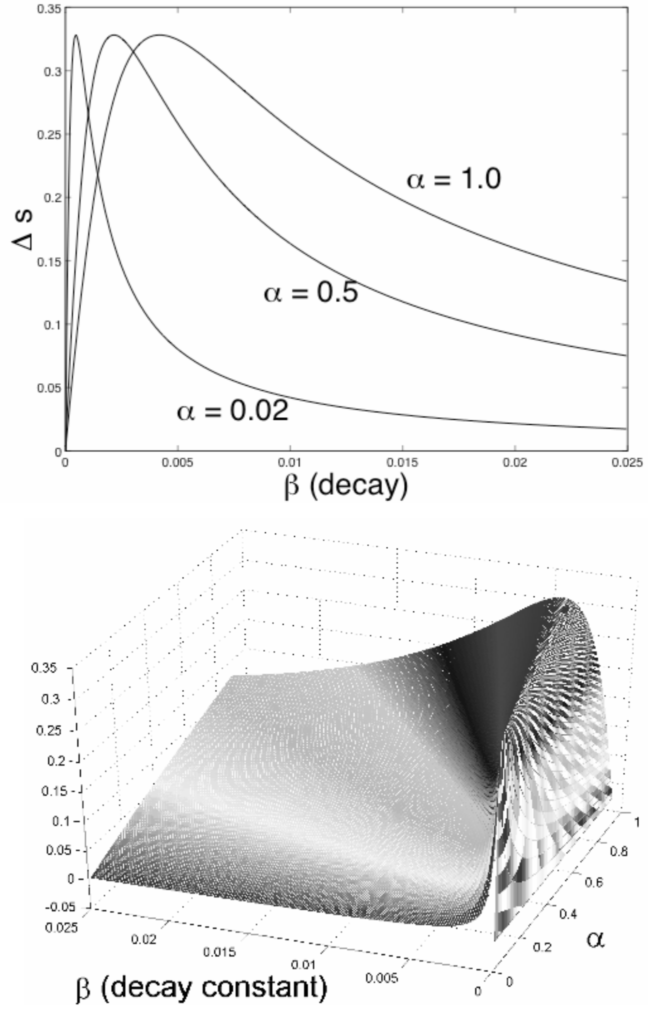


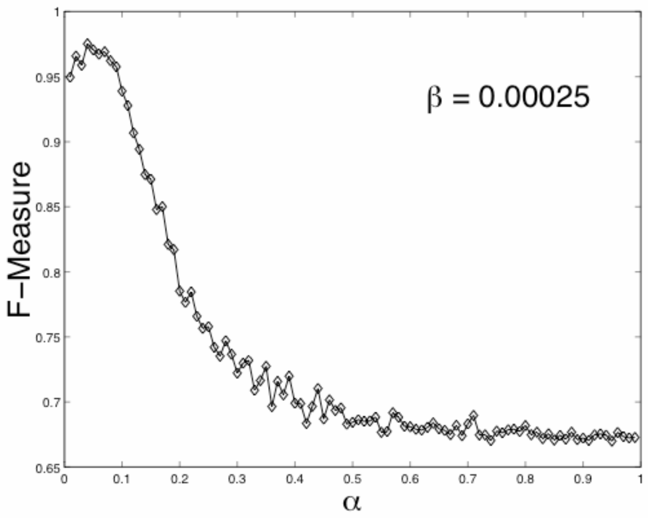
Figure 2: (top) The difference between the steady state values ΔS as a function of decay parameter β, for three different values of α; (bottom) surface plot of the ΔS(α,β).

The strong non-monotonic behavior of ΔS suggests that the classification procedure using suspicion score might be very sensitive to the model parameters (α, β). Indeed, if we assume that the separation between the averages of two distributions can be used to characterize their overlap then the classification accuracy will vary with the choice of (α, β). We examined this issue empirically in simulation. Note, that we do not take into account variation due to the subtle question of setting the correct classification threshold. Instead, to concentrate on the impact of α and β solely, we empirically find the threshold that yields the best F-Measure.

We present the result in Figure 3 where we plot the F-Measure versus α for a fixed value of the decay parameter $\beta = 0.00025$. The maximum separation of the averages for this value of β occurs at $\alpha = 0.1$, that corresponds to the value of roughly F-Measure ≈ 0.98. If one increases α, however, the detection accuracy deteriorates significantly, e.g., for $\alpha = 0.3$ the F-Measure drops below 0.75, and decreases even further as α increases.

We also note that when one changes α in the opposite direction (e.g., decreasing starting from $\alpha = 0.1$), then there is no significant deterioration in the classification accuracy. The reason for this is that while decreasing α, the score difference ΔS decreases not because two distributions are overlapping significantly, but because the values of S_c and S_b themselves are small. However, their variances are small as well so that two distributions stay separated.

Figure 3: F-Measure vs α.



4 Discussion

In this paper we presented and studied a classification model that we believe mimics guilt by association. In our model we associate a suspicion score with each individual that changes with time: it increases whenever an individual meets with another individual with non-zero score, and also decays in time so that the distribution of scores over the entire population have a well defined long term limit. Our results indicate that assigning suspicion scores, and hence the classification procedure itself, can be very sensitive to its parameters. We demonstrated this both empirically and analytically, using statistical analysis of the average score evolution in separate subpopulations.

Although in this paper we examined a single model for suspicion scoring, we think that our results might have more general implications. Indeed, despite its simplicity our model seems to capture the main elements of suspicion scoring models: increase in an individual's score due to meeting with other suspicious individuals, and decrease due to "inactivity".

More generally, it is known that only few popular data mining algorithms performed well on multiple problems without parameter adjustments (Keogh 2004). In a classification problem with suspicion scoring one usually have certain number of tunable parameters. If the behavior of the classification procedure is not robust to small adjustments in these parameters, then the accuracy of the final classification can be very poor, and inconsistent as it can change unpredictably. We conclude by noting that the results presented here emphasize the need for methods that will allow analyzing robustness of various classification algorithms with respect to their parameters.

Acknowledgments

We thank C. T. Morrison and G. W. King for the technical help in preparing the manuscript.

References

Clauset, A., M. E. J. Newman, M. E. J., and Moore, C. 2004. Finding community structure in very large networks. *Phys. Rev. E* **70**, 066111 (2004).

Dombroski, M., Fischbeck, P., and K.M. Carley, K. M. 2003. Estimating the Shape of Covert Networks. *Proc. of the 8th International Command and Control Research and Technology Symposium.*

Keogh, E., Lonardi, S., and Ratanamahatana, C. 2004. Towards Parameter-Free Data Mining. *Proc. of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* Seattle, WA.

Krebs, V. 2001. Mapping Networks of Terrorist Cells. *Connections*, 24 (3).

Macskassy, S. A. and Provost, F. J. 2003. A Simple Relational Classifier, *Workshop on Multi-Relational Data Mining in conjunction with KDD-2003*.

Neville, J. and Jensen, D. 2000. Iterative classification in relational data. *Proc. AAAI-2000 Workshop on Learning Statistical Models from Relational Data*, pages 13-20. AAAI Press, 2000.

M. E. J. Newman, M. E. J. and M. Girvan, M. 2004. Finding and evaluating community structure in networks. *Phys. Rev. E* **69**, 026113 (2004).