Title: Extraction of Social Context via Synthetic Pollination for Information Tracking and Control
Louisiana Tech University, POC: Dr. Jean Gourd

Background: We live in an age where an almost incomprehensible amount of information is being moved around at virtually the speed of light. There is no purely human way to manage the amount of information and activity taking place on a day-to-day basis. Furthermore, it is complicated by the fact that computers are perfectly capable of recording nearly every action and serving it up for human consumption. In fact, this is standard operating procedure where anything sensitive is concerned: record every transaction.

For a human being, the processing and comprehension of this mountain of data in a timely manner is intractable. Yet it is trivial to work backwards from an event and see the evidence that foreshadowed it. While forensics is important for accountability, no organization really wants to be compromised in the first place. Thus, the challenge of information security is to notice telltale signs ahead of time and react to them before the event occurs, or at the very least as the event is occurring.

The traditional approach to doing so, beyond recording everything, is to prevent access by stopping the enemy at the borders. Detection of cyber attackers is complex given the scale of modern information technology systems without considering the case of insider threats (attackers that are authorized and operating at least partially within normal parameters). This problem is staggering and is one largely of context: apparently innocent actions given appropriate context, and sometimes involving additional actors, are in fact malicious. A major approach to address cyber threats is to attempt to automatically detect usage or traffic patterns that are anomalous. But insiders are intrinsically trusted and typically have access to sensitive information and routinely access it. It only becomes noticeable when the information is divulged to a third party, which may not take place within the LAN (local area network). The job of the security professional or counter intelligence expert then becomes one of inference. From context, one must infer intent. This requires an even greater degree of analysis of the information. Furthermore, events cannot be considered solely in isolation; the proper context may only become evident when considering events separated in space and time.

Objective: This work centers around a synthetic biological mechanism that is based on the collection and analysis of network data designed to assist in the tracking and control of information flow in a LAN. This mechanism can be used to aid in the detection and mitigation of insider threats, for example, through the early prediction of adversarial behavior. Behaviorally, it generates additional context from base information in the form of social meta-data mined from interactions between users and nodes in the network. There is an ancillary benefit in that the additional context is expected to reduce the effective capabilities of the insider threat due to the increased risk of detection.

At the core of this mechanism is a concept called “pollination” that is modeled biologically after the activities of bees. In the course of utilizing a node, users will leave “pollen” indicating their interests in terms of other nodes, users, and external entities. This meta-data is distinct in that it is contextual in nature and ultimately reduces the amount of information an analyst must process while providing contextual connections that are either missing from data or difficult and expensive to correlate. Once analyzed, it can be used for early prediction of malicious behavior for the purpose of detecting insider threats early enough to assist in preventing attacks.

Approach: The proposed work centers on a mechanism called “pollination,” a distributed host-based sensor network that utilizes intelligent agents to dynamically process raw network data into meta-data. This meta-data is not designed to replace existing data, but is rather intended to provide additional context via distributed preprocessing. The goal is to measure specific social behaviors as they happen, as opposed to measuring everything and then proceeding to work backwards as is typically done in traditional forensics. This data takes the form of social context and provides information about the relationships formed between nodes and users, and allows

the tracking of information through the network.

Pollination is a scheme that relies upon the biological metaphor of bees. For the sake of the metaphor, messages within the network implementing pollination are bees while nodes are flowers. This allows for the tracking of communicating peers at the end points. This concept can be extended to infer the entire social network of a machine; however, the purpose of this work is to track and provide some measure of control of the flow of information in the network. To that end, the linking of pollen to a specific user identity within the local network is conceived. This is an additional tier of information that lies beyond the host. In essence, both the operator and the machine used are recorded. Thus, for every machine and user a set of pollen will exist at the network barrier indicating the other users and machines that form its peer group. A more thorough introduction of the pollination scheme and its potential applications can be found in our introductory paper on the topic [3].

The inclusion of host classification can allow for a less social- and more activity-oriented view of a user's communications. Additionally, pollen does not present a binary record of social interactions, and the quantity of pollen present would be indicative of the relative strength of a social connection. The pollen collected on individual messages provides for additional forensics capabilities in the form of a visit history. This visit history of individual packets could potentially be used to detect attackers or insiders trying to stealthily sniff packets even if they were not necessarily changing them. In addition, the need to pollinate incoming information at the network edge leads to the opportunity to automatically classify WAN (wide area network) traffic and thus generate a social selection of users that are interested in that traffic. This idea can also be applied to LAN traffic, with its simplest implementation being functional in that these groups of users have functional reasons to associate.

A major component of the pollination mechanism revolves around an intelligent agent-based command and control (C2) system. This component is necessary to house and display the aggregated information from the pollination network. Through the mechanism of collecting the information, data can be correlated providing for additional context and some data integrity checking. There is an inherent anomaly detection included in this correlation. Because of the double-ended nature of pollination, the reports from sender and receiver should more or less agree. Furthermore, this data can be used as a behavioral use pattern although it does not give insight into whether these patterns are normal or abnormal. However, it is likely that classification mechanisms will assist with this. The meta-data provides additional information that would normally be time-consuming to mine out of lower level data. This leads to a net reduction in the amount of data a higher level algorithm would have to consider in a first pass situation particularly with the inclusion of more sophisticated classification algorithms. Via user input and interaction, the classification mechanisms can be further modified allowing for the dynamic tweaking of the distributed sensor network.

Technically, pollination can be implemented in a number of ways, and determining the optimal method of doing so is a part of this work. The use of a manipulation of the Open System Interconnection (OSI) model [4] (in particular, the transport layer) can be leveraged for both attaching and transporting the pollen. In the OSI model, the application data to be transported is broken into packets and transmitted from source to destination. We can add additional packets by appending the pollen to the data stream, or we can manipulate the packets using packet tagging (e.g., [1, 2]). Adding additional packets can be accomplished at the application level by simply appending to the end of the intelligent agent in that data stream. The addendum is removed at the node and the pollen is recovered.

Once a comprehensive view of the system is aggregated, other techniques can be applied to the meta-data. One application is the creation of valid pollen patterns. This can be accomplished by simply keeping track of all valid types of pollen. If a packet is encountered with invalid pollen, it would be a very strong indicator of malicious behavior. In fact, it would most likely be an indicator of either an intruder who did not understand pollen or an insider attempting to cover his tracks.

The collected pollen meta-data can also act as input to other algorithms. For instance, this data provides an easily navigable map of peers that requires investigation that may not be self-evident from viewing simple log files. One potential algorithm to apply to this correlated data is to develop host-based classification to determine the normal peer-group for a machine and user. Thus, communications violating this peer group become suspicious evidence of, at the least, unofficial activities.

The aggregation of pollen measurements from each node into a centralized view is done via intelligent mobile agents. The primary purpose of these agents is to dynamically measure and monitor the state of the pollination sensors. A second purpose is to provide the ability to push information and parameters from the C2 system to the distributed framework. This information takes the form of rule sets, allowing for tweaking of the exact parameters of pollination such as growth, decay, classification, and granularity of packet history. There are also ancillary benefits to including agents; for example, return channel obfuscation. If the agent makes a decision at execution time about the return channel, it would be difficult to guess ahead of time. Thus, it would be more difficult for an insider to cover his tracks. Additionally, agents provide for the pushing of security polices or algorithms to the network. This allows for the future employment of computationally expensive algorithms on subsets of the network.

Although there are clear social context aspects of pollination, at its technical core it is a mechanism for tracking packets through the network. Thus it is closely linked to digital forensics, and one of the aspects of interest is the ability to backtrack the path of information in the network. That is, it is conceivable that, should information be leaked from the network (or even simply reach the network edge), its exact path back to the point of origin could be easily mapped via pollination (see Figure 1). This is because pollen is bidirectional: in addition to dropping off pollen, it can be picked up by a packet. Thus, for every packet we have a trace of its travels through the network. This trace would, at a minimum, provide the order of nodes visited. By combining a higher level algorithm, it would be possible to observe irregularities in the pollen pattern. There are some technical questions of the possible granularity this trace can take while not affecting performance; however, should the performance costs be significant, granularity can be varied depending upon the origin of the message. That is, packets originating from machines with more sensitive data or from machines exhibiting suspicious information may have greater granularity.

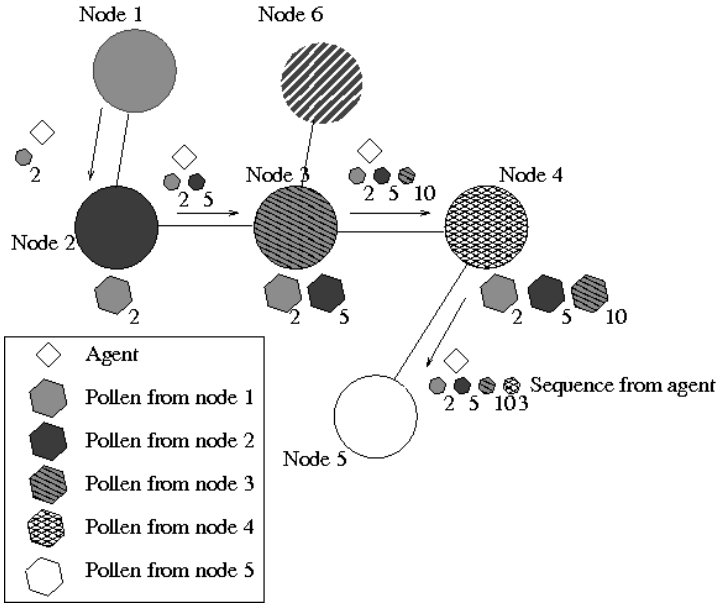

Figure 1: Path of information flow

Additionally, as an implementation detail pollen must be appended to signals within the network. Thus, upon leaving the subnet this information will be stripped to avoid confusing exterior hosts or alerting the Internet at large to the details of the pollination scheme. This stripping process must be robust and cheap because of the volume of messages it will be expected to deal with. The opportunity inherent to this chore is to allow the pollination of incoming messages in a meaningful manner. This can allow the inclusion of external site classification into the pollination record. This is a very useful aspect of pollination and factors into the constraints placed upon the stripping mechanism.

Pollen is not an all-or-nothing prospect; multiple communications will leave multiple instances of pollen. Similarly, a decay mechanism is envisioned to filter extremely sporadic communications from significance. It is important to note that decay to a zero state is assumed not to occur; thus, there will always be trace amounts of pollen for any user that has ever been communicated with. The point of allowing pollen to accumulate and decay is to provide a weighed mechanism indicating the relative strength of a social connection. This provides additional context to communications and is dynamically updated. While growth and decay mechanisms are important to the overall function of the framework, it is not necessary to determine them at the inception of the network. Any host-based implementation or hardware must be capable of performing this behavior and of being tweaked. While an initial implementation of growth and decay is an objective of this work, the ultimate determination of optimal settings will require the full system to be in place and may be one of functions of higher level components such as the C2 system or the human operator. In part, this is because the exact particulars may indeed vary depending upon the infrastructure the system is deployed on.

The uniqueness in our approach centers on the pollination mechanism. It provides socially-oriented data as opposed to traditional approaches to data organization (e.g., topographic or temporal). The key advantage of this is that it allows for the study and observation of collaborative efforts. This can directly be used to detect and mitigate insider threats in addition to controlling the flow of information. The pollination mechanism also provides for bidirectional interest tracking from both the source and destination without significant additional computational overhead or data processing. It is extremely scalable both with respect to the size of both the computing infrastructure and potential user base while also being widely deployable on a diverse variety of complex computer environments. This is largely due to the system-neutral nature of the mobile agent infrastructure. This infrastructure provides for its own security as well as a great deal of potential integration with existing and future algorithms.

Furthermore, the proposed framework is mission-neutral in that it is not tailored to detect specific or limited missions but extra-normal activities and collaborations including those with extra-organizational components. A major focus of this approach is based on social and behavioral context, ultimately providing for the integration of human factors into modeling including interactions crossing the network barrier and involving both insiders and outsiders.

The primary goal of the work in 2013 will involve developing and analyzing the optimal technical implementation for pollination. In particular, experimentation into the implementation details of the pollination scheme and its integration into a testbed network will be undertaken. There are several possibilities to implement the pollination scheme such as appending the pollen as part of the overall message or modifying the individual packets as they leave or arrive at a node. Hardware and software implementations of this will both be considered with respect to their benefit versus deployment cost. There is an additional question of what impact the mechanism will have upon network traffic; it is expected that there will be some iterative process involved in finding a minimally impacting, yet still functionally operational implementation. The performance impact of the host based aspect of pollination is also of concern, but is not expected to be as challenging to implement. Concurrently, the MAS that represents the C2 system will be initially designed. For this phase, network traffic data will be generated internally.

Future Efforts (2014-2017): Goals for 2014 include finalizing the optimal technical

implementation for pollination and the C2 system design. In addition, a suitable network testbed will be built that will support the integration of the pollination mechanism and the C2 system. This will lead to implementing the pollination prototype (generating verifiable meta-data from test data) and successfully integrating it with a prototype C2 system into the network testbed in 2015.

Goals for 2016 include optimizing the performance of the implementation for pollination within the context of the C2 system in the testbed environment. In addition, it will involve an investigation of methods to mine the meta-data for useful information that can be used to maintain information flow control. Goals for 2017 primarily involve scalability and performance related improvements. For example, the intelligent agents serve the dual purpose of managing the pollination sensor network. As human analysts make judgment calls about the relative importance of social connections or sensitivity of certain material, tweaks can be applied to the pollination network to render greater significance to those mechanisms.

In future efforts, operational testing and validation will require more specific network datasets. It is anticipated that data generated will contribute to the PREDICT dataset, a large repository of network data sources maintained by the DHS (Department of Homeland Security). The DETER testbed (also maintained by the DHS), can be utilized in later phases to deploy the mechanism in a large scale environment.

Investigators: Dr. Gourd is an Assistant Professor of Computer Science and the Program Chair of Cyber Engineering at Louisiana Tech University. He has an active research program in the areas of cyber security, distributed systems, and software engineering. His research interests include principles and methods of cyber threat avoidance, intelligent software (mobile) agents for cyber security, mobile code management and security, and cyber security education (including cyber competitions). He has worked on the development of security methods for mobile agents and has performed extensive work on developing fundamental methods to model such agents within multi-agent systems. He is particularly interested in the role intelligent agents can play in securing cyber systems. He is involved in numerous ongoing research projects with the Air Force Office of Scientific Research (AFOSR) and maintains collaborative relationships with members of industry and national research laboratories.

Dr. Gourd's Ph.D. work involved the modeling of multi-agent systems and their security characteristics. He designed the API-S Calculus, an extension to the pi-calculus intended for modeling intelligent mobile agents, including their inherent knowledge and natural grouping behavior. A long-term goal of this is to study the swarming behavior and emergent cooperative characteristics of intelligent agents within a multi-agent system in the context of cyber security.

Dr. Gourd is currently a member of the Center for Secure Cyberspace (CSC), a joint collaboration between Louisiana Tech University and Louisiana State University. Originally born in Montreal, Canada, he emigrated to the U.S. in 1983 and became a naturalized U.S. citizen in 2005. Dr. Gourd maintains an active secret government clearance.

References:

[1] A. Belenky and N. Ansari. Tracing multiple attackers with deterministic packet marking (dpm). In *PACRIM: IEEE Pacific Rim Conference on Communications, Computers and Signal Processing*, volume 1, pages 49-52, 2003.

[2] Y. Djemaiel and N. Boudriga. A global marking scheme for tracing cyber attacks. In *Proceedings of the 2007 ACM symposium on Applied Computing*, pages 170-174, 2007.

[3] J. Kackley, J. Jacobs, P. Wahjudi and J. Gourd. Pollination in maids: Detecting and combating passive intrusions in a multi-agent system. In *Proceedings of the 3rd Cyberspace Research Workshop*, pages 14-20, November 2010.

[4] D. Wetteroth. *OSI Reference Model for Telecommunications*. McGraw-Hill Professional, 2001.