Randomly Roving Agents in Wireless Sensor Networks

Hakob Aslanyan and Jose Rolim

***Abstract*—Quantitative characterization of randomly roving agents in wireless sensor networks (WSN) is studied. Below the formula simplifications, regarding the known results and publications. It is shown that the basic agent model is probabilistically equivalent to a similar simpler model and then a formula for frequencies is achieved in terms of Stirling numbers of the second kind. Stirling numbers are well studied and different estimates are known for them letting to justify the roving agents quantitative characteristics.**

***Index Terms*—Intrusion detection system, wireless sensor network, roving agents.**

I. Introduction

This work, inspired by [1], [2], [3] considers roving agents' numerical characterization in ad-hoc pervasive and trustworthy networks. Agents are autonomous, moving, and intelligent software structures capable to play a sensitive role in advanced monitoring, computation and protection systems. Intrusion detection systems (IDS) based on roving agents [2] are addressed particularly. They appear as complementary mean to the ordinary cryptographic protection tools of computers and networks. Such IDS use software agent based monitoring and data collection, watching the inside processes of a computer, registering LOG files of application software systems, sniffing and recording communication protocols. Watching the whole network behavior they are better suited to warn approaching attacks and malfunctioning. Data mining agents (DMA) and Data fusion agents (DFA) are examples of information integration tools in networks [3]. In large networks, moreover when its structure is not predefined such as wireless sensor networks [1] it is natural to consider independent, randomly roving agents, requiring that they are able to collect enough information in total, mining the necessary knowledge about the intrusion. This framework is studied in [3], which prove formulas for the number of DMA sufficient to monitor the given size areas of networks. The formula received is complex and impractical because of their use of nested sums by different parameters. By the same reason [3] considers software simulations to understand the typical number of agents required for monitoring a given size networks. Our work tends to prove simple estimates for the same numerical characteristics of WSN analytically.

II. Roving Agents Model

DMA roams around randomly in a network and acquires environmental information. It is lightweight using simplest mining algorithms. DFA is for integration of DMA set actions. DFA may act as an intrusion detection tool and then its power depends on information collected by DMA in network.

Let we are given a network N of n nodes $v_1, v_2, \ldots, v_n$. Some fixed amount of information $\Im_i$ is allocated at node v_i. There are k DMA $a_1, a_2, \ldots, a_k$. Each agent visits exactly m different nodes and obtains the unique information content in each such node. DMA pass all collected information to DFA. Denote by $P_k(n,m,t)$ the probability that DFA contains exactly t information blocks of network nodes when k agents randomly visit m of n nodes each. The formula for $P_k(n,m,t)$ proven in [3] looks as:

$$\begin{aligned} & P_k(n,m,t) = \\ & = \binom{n}{m}^{-(k-1)} \sum_{m_2, m_3, \ldots, m_{k-1}=0}^{m} \binom{m}{m_2} \binom{n-m}{m-m_2} \\ & \binom{2m-m_2}{m_3} \binom{n-2m+m_2}{m-m_3} \\ & \binom{(k-2)m-m_2-\ldots-m_{k-2}}{m_{k-1}} \\ & \binom{n-(k-2)m+m_2+m_{k-2}}{m-m_{k-1}} \\ & \binom{(k-1)m-m_2-\ldots-m_{k-1}}{km-t-m_2-\ldots-m_{k-1}} \\ & \binom{n-(k-1)m+m_2+m_{k-1}}{t-(k-1)m+m_2+\ldots+m_{k-1}}, k \geq 4 \end{aligned} \tag{1}$$

Formulas for smaller k given in [3] look similar to (1). Of course these formulas are unobservable and simplifications or approximations are of interest. By this same reason [3] proves formulas, considering computer simulation, to understand the typical numbers of agents necessary to retrieve the required information in network. Modification of "exactly m" condition in agent distribution scheme is also important and will be considered.

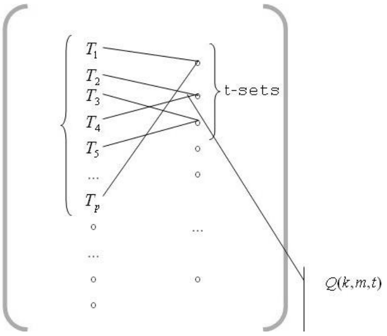

Fig. 1. Agent sets distribution in terms of trials and node sets. Left column contains outcomes of k by m trials (each T_i is an ordered collection of k m-subsets). Right column contains all the subsets of node set N.

Manuscript received November 28, 2012; revised March 20, 2013.

The authors are with Theoretical Computer Science group at University of Geneva (e-mail: hakob.aslanyan@unige.ch).

DOI: 10.7763/IJCTE.2013.V5.789

III. Coverage Characterization of Roving Agents

Let we are given a set of nodes $N = \{v_1,...,v_n\}$ and $S_1,...,S_k$ are k arbitrary subsets of N, each of size m. S_i corresponds to the set of nodes visited by agent i. We consider a probability distribution scheme over N, and suppose that m-subsets S_i are equiprobable and independent in this scheme. Having in total C_n^m m-subsets the probability of one of them is equal to $1/C_n^m$. We are interested in knowing the probabilistic characteristics of the union $\bigcup_{i=1}^k S_i$ and its size $t = \left|\bigcup_{i=1}^k S_i\right|$. In particular, what is the probability that union of those subsets contains exactly t elements?

$$p_k(n,m,t) = \Pr\left(\left|\bigcup_{i=1}^{k} S_i\right| = t\right) \tag{2}$$

To a collection of subsets $S_1,...,S_k$ of N nodes corresponds a matrix $A^{kxn} = \{a_{ij}\}$ where

$$a_{ij} = \begin{cases} 1 & \text{if } v_j \in S_i \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

From the fact that each S_i contains exactly m elements follows that each row of matrix A contains exactly m 1 s and $n-m$ 0 s. If $\left|\bigcup_{i=1}^{k} S_i\right| = t$ then there are t columns of A which contain at least one 1 and $n-t$ columns which do not contain 1. The number of $k \times n$ matrixes with m ones on each row and with exactly $n-t$ columns with no 1 is $C_n^t \cdot Q(k,m,t)$ where $Q(k,m,t)$ is the number of $k \times t$ matrixes with m ones on each row and at least one 1 on each column.

Alternatively, consider the following schematic presentation of roving agents' distribution. Left column vertices in the scheme presented in Fig. 1 contain all the arrangements $T_1, T_2, ...$ of k agents (ordered collections of k m-node-subsets). From combinatorial perspective agents and nodes are distinguishable but m-node subsets are considered as usual sets - different elements and no ordering. Total number of such arrangements is equal to $\left(C_n^m\right)^k$. Part of these arrangements covers exactly t nodes and let that vertexes corresponding to this arrangements are $T_1, T_2, ..., T_p$. In this notation p is the unknown number that we want to compute. Right side column vertices correspond to all subsets of node set N and part of these sets are of size t. In principle, node subset sizes may vary from 0 to n but in our experiment it may take values from m to $\min(km,n)$.

We draw an edge between an arrangement (left column) and a node subset (right column) if node subset is covered by that arrangement. Each arrangement is incident to exactly one edge (and subset). Each t-subset appears in different arrangements and this number is common for all t-subsets and is given by $Q(k,m,t)$.

$Q(k,m,t)$ can be calculated by inclusion-exclusion principle. We use the matrix model for calculating $Q(k,m,t)$. First, over a $k \times t$ matrix we take the whole set of unconstrained arrangements i.e. all matrices with m 1s on rows, then we remove from this all the arrangements where at least one column is initially filled with 0 (such matrices do not obey to conditions we require), then add arrangements with at least 2 empty columns, etc. The formula representation of related quantities is:

$$\begin{aligned} Q(k,m,t) &= \left(C_t^m\right)^k - C_t^1 \cdot (C_{t-1}^m)^k + C_t^2 \cdot (C_{t-2}^m)^k - \dots \\ &+ (-1)^{t-m} (C_t^{t-m})^k = \\ &\sum_{i=0}^{t-m} (-1)^i C_t^i \cdot \left(C_{t-i}^m\right)^k \end{aligned} \tag{4}$$

We have proven

Theorem 1.

$$P_k(n,m,t) = \frac{C_n^t \cdot \sum_{i=0}^{t-m} (-1)^i C_t^i \cdot \left(C_{t-i}^m\right)^k}{\left(C_n^m\right)^k} \tag{5}$$

First of all here we receive a real simplification of (1). The formula received is still complex, but it might be approximated and the applied Markov inequality may give asymptotic estimates of t-subset probabilities [4].

Another important characteristic, the mean value of subset size t, might be computed as:

$$\begin{aligned} &\sum_{t=m}^{\min(km,n)} t \cdot P_k(n,m,t) = \\ &\sum_{t=m}^{\min(km,n)} \frac{t \cdot C_n^t \cdot \sum_{i=0}^{t-m} (-1)^i C_t^i \cdot \left(C_{t-i}^m\right)^k}{\left(C_n^m\right)^k} \end{aligned} \tag{6}$$

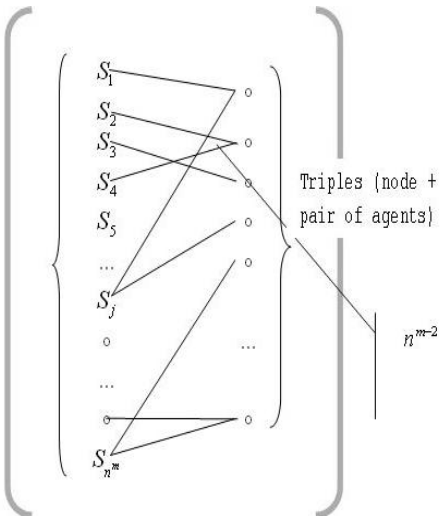

Fig. 2. Agents distribution on WSN node sets. Left column contains outcomes of m trials (each S_i is an ordered collection of m nodes), right column contains triples, node and two different agents

IV. On Node Repetition Limitations in an Roving Agent Scheme

Let us consider the scene of random distribution of m agents over the n WSN nodes (here we do not consider k agents but m agents, and each individual agent visits exactly one node). Agents are dropped over the node set one by one, independently, and with equal probabilities for nodes. Allocating all m agents we receive a collection of nodes visited by agents, probably with multiple agents that visited the same node.

The total number of different allocations is n^m. Among them are 1 node allocations (all the agents visit the same node), their number is n, 2 node allocations, they are $C_n^2(2^m - 2)$ and so on. The largest are m node allocations (m-sets), when agents are distributed in all different nodes, the number of such allocations $n(n-1)\ldots(n-m+1)$. We are interested in the frequencies of allocation sizes when at least 2 agents are allocated at the same node (sizes from 1 to $m-1$), or complementary, the share of allocations with all different nodes.

One of the classical approaches of determining typical cases in distributions is when Markov or Chebyshev inequality is applied. In this way we consider a scheme presented in Fig. 2 similar to one presented in Fig. 1 to compute the mean value of the number of allocated nodes in random distribution of m agents (of memory 1) over n WSN nodes.

Thus, the number of right side vertices in the scheme, where each vertex is a triple; a node and a pair of agents, is nC_m^2. Edges are connecting an allocation (from left column) to a node with a given pair of agents that visited that node (right column). We compute the average number of edges $M(\upsilon_{n,m})$ incident to each allocation as

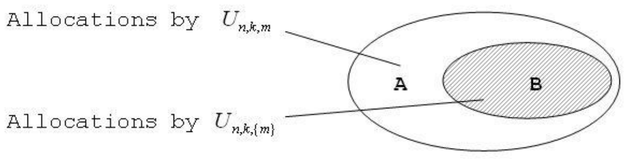

Fig. 3. Allocations by $U_{n,k,\{m\}}$ and $U_{n,k,m}$

$$M(\upsilon_{n,m}) = \frac{nC_m^2 n^{m-2}}{n^m} = \frac{C_m^2}{n} \tag{7}$$

Applying Markov inequality $\Pr\{\upsilon_{n,m} \geq \varepsilon\} \leq M(\upsilon_{n.m})/\varepsilon$ and taking $\varepsilon = 1$, we get an upper bound C_m^2/n for the probability of agents repeating on the nodes . If $C_m^2/n \to 0$ with $n, m \to \infty$, then we receive that almost all allocations consist of all different agents at nodes.

V. Comparison of Agent Allocation Schemes

In this point we will define and consider two basic probability distributions tightly related to each other.

1) First distribution $U_{n,k,\{m\}}$ is composed of k independent consecutive allocations of m-node subsets over the WSN area of n nodes. $(C_n^m)^k$ Outcomes of trials are ordered collections of m-subsets of WSN nodes. These collections may cover all node subsets of sizes from m to $\min(km, n)$.

2) Second distribution scheme $U_{n,k,m}$, which we consider and compare with the basic distribution $U_{n,k,\{m\}}$ considered above, consists of k consecutive and independent stages; each stage allocates m elements consecutively and independently over the WSN area of n nodes. Outcomes of these trials are all n^{km} ordered collections of nodes. These collections may cover all node subsets of sizes from 1 to $\min(km, n)$.

In one individual stage of $U_{n,k,m}$ we have $m!$ orderings of a single allocation of m-subset of one step of $U_{n,k,\{m\}}$. This is to be taken into account comparing the schemes $U_{n,k,\{m\}}$ and $U_{n,k,m}$. This difference can also be seen comparing the one stage outcomes of $U_{n,k,\{m\}}$ and $U_{n,k,m}$. Represent C_n^m of model $U_{n,k,\{m\}}$ as

$$\frac{n!}{m!(n-m)!} = \frac{n(n-1)\ldots(n-m+1)}{m!} \tag{8}$$

Numerator of the last ratio is the counterpart of n^m of $U_{n,k,\{m\}}$ model, and $m!$ is the coefficient we mentioned about. Comparing $U_{n,k,\{m\}}$ and $U_{n,k,m}$, first we note that outcomes of $U_{n,k,\{m\}}$ are part of outcomes of $U_{n,k,m}$ and hence they may have higher probabilities. Consider the probability p_j of an event, that in stage j of $U_{n,k,m}$, all the allocated m elements are different. Then $P = p_1 \cdot p_2 \cdot \ldots \cdot p_k$ is the probability that in all k stages allocated m elements are different. In different stages allocations of course may intersect. Outcomes of $U_{n,k,\{m\}}$ multiplied with this probabilities are equal to probabilities of $U_{n,k,m}$, part B of intersection of outcomes in Fig. 3. p_j was estimated in previous point as a value tending to 1 asymptotically. We may extend this proposition to the entire value P. Formally we use the property that probability of union of events is less or equal the sum of event probabilities:

$$\begin{aligned} &\Pr\{(\upsilon_{n,m} \geq \varepsilon | q=1) \vee \ldots \vee (\upsilon_{n,m} \geq \varepsilon | q=k)\} \leq \\ &\leq k \cdot \Pr\{\upsilon_{n,m} \geq \varepsilon\} \leq \frac{k \cdot M(\upsilon_{n.m})}{\varepsilon} \end{aligned} \tag{9}$$

Then $kC_m^2/n \to 0$ with $n, m, k \to \infty$ is a sufficient condition (upper estimate) for repetition probability to tend to zero. The sufficient condition $km^2/n \to 0$ for allocation of all m agents in all k consecutive stages to different nodes is naturally acceptable in WSN which have a very large nodes set as a rule. The final picture is

1) Part B allocations (Fig. 3) appear in $U_{n,k,m}$ with probability P tending to 1.

2) Relative probability distribution among the elements of B is identical in both models $U_{n,k,\{m\}}$ and $U_{n,k,m}$
3) Event probability in model $U_{n,k,\{m\}}$ is not less than in $U_{n,k,m}$ multiplied by P.
4) Probabilities of t-subset allocations under the model $U_{n,k,m}$ have formulas similar to the ones for model $U_{n,k,\{m\}}$ considered above.

If $R(k,m,t)$ denotes the number of t-node allocations in model $U_{n,k,m}$ then the formal representation of $R(k,m,t)$ is similar to the formula for $Q(k,m,t)$. Considered above can be achieved by the same inclusion exclusion method:

$$R(k,m,t) = t^{mk} - C_t^1 \cdot (t-1)^{mk} + C_t^2 \cdot (t-2)^{mk} - \ldots$$
$$\ldots + (-1)^{t-1}(t-1)^{mk} =$$
$$= \sum_{i=0}^{t-1} (-1)^i C_t^i \cdot (t-i)^{mk} \tag{10}$$

On this basis we formulate.

Theorem 2. If $kC_m^2/n \to 0$ with $n,m,k \to \infty$, then t-node allocation probabilities in models $U_{n,k,\{m\}}$ and $U_{n,k,m}$ have the following relation

$$\frac{C_n^t Q(k,m,t)}{(C_n^m)^k} \cdot P \le \frac{C_n^t R(k,m,t)}{n^{km}}$$

Finally, we note that $R(k,m,t)$ has equivalent presentation in terms of Stirling numbers of the second kind ([5])

$$S(N,K) = \frac{1}{K!} \sum_{j=0}^{K} (-1)^j C_K^j (K-j)^N \tag{11}$$

Here we used the fact that allocation of k consecutive and independent stages of m elements over the WSN area of n nodes is equivalent to allocation of km elements over that area. Note that the difference between the formulas for $Q(k,m,t)$ and $R(k,m,t)$ is the summation limits. In case of $R(k,m,t)$ formally we may add the zero term for $i=t$, and then we receive

$$R(k,m,t) = t!S(mk,t) \tag{12}$$

which is the final postulation of this paper.

VI. Conclusion

WSN and software agent systems are important application technique for many areas. Being hard algorithmically and complex in model level these systems require special economy regimes to work. Above we considered an intrusion detection system based on roving agents and achieved simple formulas that allow to understand the number of agents required for monitoring a given size network. One of the main formulas is given in terms of combinatorial Stirling numbers and known asymptotic estimates for them [5] allows to adopt the monitoring regime in an optimal way.

References

[1] U. Brandes and T. Erlebach, *Network Analysis–Methodological Foundations*, Springer-Verlag, Berlin Heidelberg, 2005.

[2] C. Krugel, T. Toth, and E. Kirda, “A mobile agent based intrusion detection system,” in *Proc. First International IFIP TC-11 WG 11.4 Working Conference on Network Security*, 2001.

[3] I. S. Moskowitz, M. H. Kang, L. W. Chang, and E. Garth, “Longdon. Randomly roving agents for intrusion detection,” Technical report, Naval research laboratory, Washington D.C., 2001.

[4] Y. I. Medvedev and G. I. Ivchenko, “Asimptotical expansions of finite differences of power function in an arbitrary point,” *Theory of Probability and Applications*, vol. 10, pp. 151–156, 1965.

[5] R. Chelluri, L. B. Richmond, and N. M. Temme, “Asymptotic estimates for generalized stirling numbers,” *Report-Modelling, Analysis and Simulation, CWI*, Amsterdam, Netherlands, 1997.

Hakob Aslanyan received his bachelor and masters degrees at Yerevan State University, Yerevan Armenia. Currently he is a PhD student at Theoretical Computer Science group at University of Geneva under the direction of Prof. J. Rolim. His main topics of interest are combinatorial optimization, graph theory, approximation and exact algorithms, hardness of approximation, network design and connectivity

Jose Rolim is full professor and head of the TCS-sensor lab of the Computer Science Department of the University of Geneva. He received his PhD from the Computer Science Department of the University of California at Los Angeles under the direction of Prof. Sheila Greibach. His main topics of interest are Theoretical Computer Science (TCS), Algorithmic aspects of ad-hoc networks, computing with wireless sensor networks.