A Probabilistic Bound on the Basic Role Mining Problem and Its Applications

Alessandro Colantonio[1], Roberto Di Pietro[2], Alberto Ocello[3], and Nino Vincenzo Verde[4]

[1] Engiweb Security, Roma, Italy
and Università di Roma Tre, Roma, Italy
alessandro.colantonio@eng.it,
colanton@mat.uniroma3.it
[2] Università di Roma Tre, Roma, Italy and UNESCO Chair in Data Privacy, Tarragona, Spain
dipietro@{mat.uniroma3.it,urv.cat}
[3] Engiweb Security, Roma, Italy
alberto.ocello@eng.it
[4] Università di Roma Tre, Roma, Italy
nverde@mat.uniroma3.it

Abstract. The aim of this paper is to describe a new probabilistic approach to the role engineering process for RBAC. We address the issue of minimizing the number of roles, problem known in literature as the Basic Role Mining Problem (*basicRMP*). We leverage the equivalence of the above issue with the vertex coloring problem. Our main result is to prove that the minimum number of roles is sharply concentrated around its expected value. A further contribution is to show how this result can be applied as a stop condition when striving to find out an approximation for the *basicRMP*. The proposal can be also used to decide whether it is advisable to undertake the efforts to renew a RBAC state. Both these applications can result in a substantial saving of resources. A thorough analysis using advanced probabilistic tools supports our results. Finally, further relevant research directions are highlighted.

1 Introduction

An *access control model* is an abstract representation of security technology, providing a high-level logical view to describe all peculiarities and behaviors of an access control system. The *Role-Based Access Control* (RBAC, [1]) is certainly the most widespread access control model proposed in the literature for medium to large-size organizations. The simplicity of this model is one of the main reasons for its adoption: a role is just a collection of privileges, while users are assigned to roles based on duties to fulfil [10].

The migration to RBAC introduces several benefits, such as simplified system administration, enhanced organizational productivity, reduction in new employee downtime, enhanced system security and integrity, simplified regulatory compliance, and enhanced security policy enforcement [6]. To maximize all these advantages, the model must be customized to describe the organizational roles and functions [3]. However, this migration process often has a high economic impact. To optimize the customization, the *role*

D. Gritzalis and J. Lopez (Eds.): SEC 2009, IFIP AICT 297, pp. 376–386, 2009.

engineering discipline has been introduced. It can be defined as the set of methodologies and tools to define roles and to assign permissions to roles according to the actual needs of the company [5].

To date, various role engineering approaches have been proposed in order to address this problem. They are usually classified in literature as: *top-down* and *bottom-up*. The former carefully decomposes business processes into elementary components, identifying which system features are necessary to carry out specific tasks. This approach is mainly manual, as it requires a high level analysis of the business. The bottom-up class searches legacy access control systems to find *de facto* roles embedded in existing permissions. This process can be automated resorting to data mining techniques, thus leading to what is usually referred to as *role mining*.

Since the bottom-up approach can be automated, it has attracted a lot of interest from researchers who proposed new data mining techniques particularly designed for role engineering purposes. Various role mining approaches can be found in the literature [17, 3, 20, 18, 19, 22, 7, 12, 16]. A problem partially addressed in these works is the "interestingness" of roles. Indeed, the importance of role completeness and role management efficiency resulting from the role engineering process has been evident from the earliest papers on the subject. However, only recently have researchers started to formalize the role-set optimality concept. One possible optimization approach is minimizing the total number of roles [18, 7, 12]. Yet, the identification of the role-set that describes the access control configuration with the minimum number of roles is an NP-complete problem [18]. Thus, all of the aforementioned papers just offer an approximation of the optimal solution in order to address the complexity of the problem. However, since none of them quantify the introduced approximations, it is not possible to estimate the quality of the proposed role mining algorithm outcomes.

Contributions. In this paper we provide a probabilistic method to optimize the number of roles needed to cover all the existing user-permission assignments. The method leverages a known reduction of the role number minimization problem to the chromatic number of a graph. The main contribution of this work is to prove that the optimal role number is sharply concentrated around its expected value. We further show how this result can be used as a *stop condition* when striving to find an approximation of the optimum for any role mining algorithm. The corresponding rational is that if a result is close to the optimum, and the effort required to discover a better result is high, it might be appropriate to accept the current result.

Roadmap. This paper is organized as follows: Section 2 reports relevant related works. Section 3 summarizes the main concepts used in the rest of the paper; namely, a formal description of the RBAC model, some probabilistic tools, and a brief review of graph theory. In Section 4 the role minimization problem is formally described. Section 5 provides the main theoretical result and discusses some practical applications of this result. Finally, Section 6 presents some concluding remarks and further research directions.

2 Related Work

Kuhlmann et al. [11] first introduced the term "role mining", trying to apply existing data mining techniques (i.e., clustering similar to k-means) to implement a bottom-up

approach. The first algorithm explicitly designed for role engineering is described in [17], applying hierarchical clustering on permissions. Another example of a role mining algorithm is provided by Vaidya et al. [20]; they applied subset enumeration techniques to generate a set of candidate roles, computing all possible intersections among permissions possessed by users.

The work of Colantonio et al. [3,4] represents the first attempt to discover roles with business meanings. The authors define a metric for evaluating good collections of roles that can be used to minimize the number of candidate roles. Vaidya et al. [18,19] also studied the problem of finding the minimum number of roles covering all permissions possessed by the users, calling it the basic *Role Mining Problem* (*basicRMP*). They also demonstrated that such a problem is NP-complete. Ene et al. [7] offer yet another alternative model to minimize the number of candidate roles. In particular, they reduced the problem to the well-known minimum clique partition problem or, equivalently, to the minimum biclique covering. Actually, not only is the role number minimization equivalent to the clique covering, but it has been reduced to many other NP problems, like binary matrices factorization [12] and tiling database [9] to cite a few. These reductions make it possible to apply fast graph reduction algorithms to exactly identify the optimal solution for some realistic data set—however, the general problem is still NP-complete.

Recently, Frank et al. [8] proposed a probabilistic model for RBAC. They defined a framework that expresses user-permission relationships in a general way, specifying the related probability. Through this probability it is possible to elicit the role-user and role-permission assignments which then make the corresponding direct user-permission assignments more likely. The authors also presented a sampling algorithm that can be used to infer their model parameters. The algorithm converges asymptotically to the optimal value; the approach described in this paper can be used to offer a stop condition for the quest to the optimum.

3 Background

In this section we review all the notions used in rest of the paper, namely the RBAC entities, some probabilistic tools, and some graph theory concepts.

3.1 Role-Based Access Control

The RBAC entities of interest are:

- $PERMS$, the set of access permissions;
- $USERS$, the set of all system users;
- $ROLES$, the set of all roles, namely permission combinations.
- $UA \subseteq USERS \times ROLES$, the set of user-role assignments; given a role, the function assigned_users: $ROLES \rightarrow 2^{USERS}$ identifies all the assigned users.
- $PA \subseteq PERMS \times ROLES$, the set of permission-role assignments; given a role, the function assigned_perms: $ROLES \rightarrow 2^{PERMS}$ identifies all the assigned perms.

In addition to the RBAC standard entities, the set $UP \subseteq USERS \times PERMS$ identifies permission to user assignments. In an access control system it is represented by entities describing access rights (e.g., access control lists).

3.2 Martingales and Azuma-Hoeffding Inequality

We shall now present some definitions and theorems that provide the mathematical basis we will further discuss later on in this paper. In particular, we introduce: martingales, Doob martingales, and the Azuma-Hoeffding inequality. These are well known tools for the analysis of randomized algorithms [15, 21].

Definition 1 (Martingale). *A sequence of random variables $Z_0, Z_1, \ldots, Z_n$ is a* martingale with respect to the sequence $X_0, X_1, \ldots, X_n$ *if for all $n \geq 0$, the following conditions hold:*

- *Z_n is function of $X_0, X_1, \ldots, X_n$,*
- *$\mathbb{E}[|Z_n|] \leq \infty$,*
- *$\mathbb{E}[Z_{n+1} \mid X_0, \ldots, X_n] = Z_n$,*

where the operator $\mathbb{E}[\cdot]$ indicates the expected value of a random variable. A sequence of random variables $Z_0, Z_1, \ldots$ is called martingale *when it is a martingale with respect to himself. That is $\mathbb{E}[|Z_n|] \leq \infty$ and $\mathbb{E}[Z_{n+1} \mid Z_0, \ldots, Z_n] = Z_n$.*

Definition 2 (Doob Martingale). *A* Doob martingale *refers to a martingale constructed using the following general approach. Let $X_0, X_1, \ldots, X_n$ be a sequence of random variables, and let Y be a random variable with $\mathbb{E}[|Y|] < \infty$. (Generally Y, will depend on $X_0, X_1, \ldots, X_n$.) Then*

$$Z_i = \mathbb{E}[Y \mid X_0, \ldots, X_i], \;\; i = 0, 1, \ldots, n,$$

gives a martingale with respect to $X_0, X_1, \ldots, X_n$.

The previous construction assures that the resulting sequence $Z_0, Z_1, \ldots, Z_n$ is always a martingale.

A useful property of the martingales that we will use in this paper is the Azuma-Hoeffding inequality [15]:

Theorem 1 (Azuma-Hoeffding inequality). *Let $X_0, \ldots, X_n$ be a martingale s.t.*

$$B_k \leq X_k - X_{k-1} \leq B_k + d_k,$$

for some constants d_k and for some random variables B_k that may be functions of $X_0, X_1, \ldots, X_{k-1}$. Then, for all $t \geq 0$ and any $\lambda > 0$,

$$\Pr(|X_t - X_0| \geq \lambda) \leq 2 \exp\left(\frac{-2\lambda^2}{\sum_{k=1}^{t} d_k^2}\right). \tag{1}$$

The Azuma-Hoeffding inequality applied to the Doob martingale gives the so called *Method of Bounded Differences* (MOBD) [14].

3.3 Graphs Modeling

This section describes some graph related concepts that will be used to generate our model. A *graph* G is an ordered pair $G = \langle V, E \rangle$, where V is the set of vertices, and E

is a set of unordered pairs of vertices. We say that $v, w \in V$ are *endpoints* of the edge $\langle v, w \rangle \in E$. Given a subset S of the vertices $V(G)$, then the subgraph *induced* by S is the graph where the set of vertices is S, and the edges are the members of $E(G)$ such that the corresponding endpoints are both in S. We denote with $G[S]$ the subgraph induced by S. A *bipartite graph* is a graph where the set of vertex can be partitioned into two subsets V_1 and V_2 such that $\forall \langle v_1, v_2 \rangle \in E(G), v_1 \in V_1, v_2 \in V_2$.

A *clique* is a subset S of vertices in G, such that the subgraph induced by S is a complete graph, namely for every two vertices in S there exists an edge connecting the two. A *biclique* in a bipartite graph, also called *bipartite clique*, is a set of vertices $B_1 \subseteq V_1$ and $B_2 \subseteq V_2$ such that $\langle b_1, b_2 \rangle \in E$ for all $b_1 \in B_1$ and $b_2 \in B_2$. In other words, if G is a bipartite graph, a set S of vertices $V(G)$ is a biclique if and only if the subgraph induced by S is a complete bipartite graph. In this case we will say that the vertices of S induce a biclique in G. A *maximal* clique or biclique is a set of vertices that induces a complete subgraph, and that is not a subset of the vertices of any larger complete subgraph.

A *clique cover* of G is a collection of cliques $C_1, \ldots, C_k$, such that for each edge $\langle u, v \rangle \in E$ there is some C_i that contains both u and v. A *minimum clique partition* (MCP) of a graph is a smallest by cardinality collection of cliques such that each vertex is a member of exactly one of the cliques; it is a partition of the vertices into cliques. Similar to the clique cover, a *biclique cover* of G is a collection of biclique $B_1, \ldots, B_k$ such that for each edge $\langle u, v \rangle \in E$ there is some B_i that contains both u and v. We say that B_i covers $\langle u, v \rangle$ if B_i contains both u and v. Thus, in a biclique cover, each edge of G is covered at least by one biclique. A *minimum biclique cover* (MBC) is the smallest collection of bicliques that covers the edges of a given bipartite graph, or in other words, is a biclique cover of minimum cardinality.

4 Problem Modelling

4.1 Definitions

The following definitions are required to formally describe the problem:

Definition 3 (System Configuration). *Given an access control system, we refer to its* configuration *as the tuple* $\varphi = \langle USERS, PERMS, UP \rangle$*, that is the set of all existing users, permissions, and the corresponding relationships between them.*

A system configuration represents the user authorization state before migrating to RBAC, or the authorizations derivable from the current RBAC implementation—in this case, the user-permission relationships may be derived as:

$$UP = \{\langle u, p \rangle \mid \exists r \in ROLES : u \in \text{assigned\_users}(r) \ \wedge \ p \in \text{assigned\_perms}(r)\}$$

Definition 4 (RBAC State). *An RBAC* state *is a tuple* $\psi = \langle ROLES, UA, PA \rangle$*, namely an instance of all the sets characterizing the RBAC model.*

An RBAC state is used to obtain a system configuration. Indeed, the role engineering goal is to find the "best" state that correctly describes a given configuration. In particular, we are interested in finding the following kind of states:

Definition 5 (Candidate Role-Set). *Given an access control system configuration φ, a* candidate role-set *is the RBAC state ψ that "covers" all possible combinations of permissions possessed by users according to φ, namely a set of roles such that the union of related permissions exactly matches with the permissions possessed by the user. Formally*

$$\forall u \in USERS, \exists R \subseteq ROLES : \bigcup_{r \in R} \text{assigned\_perms}(r) = \{p \in PERMS \mid \langle u, p \rangle \in UP\}.$$

Definition 6 (Cost Function). *Let Φ, Ψ be respectively the set of all possible system configurations and RBAC states. We refer to the* cost function *cost as*

$$cost\colon \Phi \times \Psi \to \mathbb{R}^+$$

where $\mathbb{R}^+$ indicates positive real numbers including 0; it represents an administration cost estimate for the state ψ used to obtain the configuration φ.

The administration cost concept was first introduced in [3]. Leveraging the cost metric enables to find candidate role-sets with the lowest effort to administer them.

Definition 7 (Optimal Candidate Role-Set). *Given a configuration φ, an* optimal candidate role-set *is the corresponding configuration ψ that simultaneously represents a candidate role-set for φ and minimized the cost function $cost(\varphi, \psi)$.*

The main goal related to mining roles is to find optimal candidate role-sets. In the next section we focus on optimizing a particular cost function. Let *cost* indicate the number of needed roles. The role mining objective then becomes to find a candidate role-set that has the minimum number of roles for a given system configuration. This is exactly the *basicRMP*. We will show that this problem is equivalent to that of finding the chromatic number of a given graph. Using this problem equivalence, we will identify a useful property on the concentration of the optimal candidate role-sets. This allows us to provide a stop condition for any iterative role mining algorithm that approximates the minimum number of roles.

4.2 The Proposed Model

Given the configuration $\varphi = \langle USERS, PERMS, UP \rangle$ we can build a bipartite graph $G = \langle V, E \rangle$, where the vertex set V is partitioned into the two disjoint subset *USERS* and *PERMS*, and where E is a set of pairs $\langle u, p \rangle$ such that $u \in USERS$ and $p \in PERMS$. Two vertices u and p are connected if and only if $\langle u, p \rangle \in UP$.

A biclique coverage of the graph G identifies a unique candidate role-set for the configuration φ [7], that is $\psi = \langle ROLES, UA, PA \rangle$. Indeed, every biclique identifies a role, and the vertices of the biclique identify the users and the permission assigned to this role. Let the function *cost* return the number of roles, that is:

$$cost(\varphi, \psi) = |ROLES| \tag{2}$$

In this case, minimizing the cost function is equivalent to finding a candidate role-set that minimizes the number of roles. This corresponds to *basicRMP*. Let $\mathscr{B}$ a biclique coverage of a graph G, we define the function $cost'$ as:

$$cost'(\mathscr{B}) = cost(\varphi, \psi)$$

where ψ is the state $\langle UA, PA, ROLES \rangle$ that can be deduced by the biclique coverage $\mathscr{B}$ of G, and G is the bipartite graph built from the configuration φ that is uniquely identified by $\langle USERS, PERMS, UP \rangle$. In this model, the problem of finding an optimal candidate role-set can be equivalently expressed as finding a biclique coverage for a given bipartite graph G that minimizes the number of required bicliques. This is exactly the *minimum biclique coverage* (MBC) problem. In the following we first recall both the reduction of the MBC problem to the *minimum clique partition* (MCP) problem [7] and the reduction of MCP to the chromatic number problem.

From the graph G, it is possible to construct a new undirected unipartite graph G' where the edges of G become the vertices of G': two vertices in G' are connected by an edge if and only if the endpoints of the corresponding edges of G induce a biclique in G. Formally:

$$G' = \left\langle E,\ \{\langle e_1, e_2 \rangle \mid e_1, e_2 \text{ induce a biclique in } G\} \right\rangle$$

The vertices of a (maximal) clique in G' correspond to a set of edges of G, where the endpoints induce a (maximal) biclique in G. The edges covered by a (maximal) biclique of G induce a (maximal) clique in G'. Thus, every biclique edge cover of G corresponds to a collection of cliques of G' such that their union contains all of the vertices of G'. From such a collection, a clique partition of G' can be obtained by removing any redundantly covered vertex from all but one of the cliques to which it belongs to. Similarly, any clique partition of G' corresponds to a biclique cover of G. Thus, the size of a minimum biclique coverage of a bipartite graph G is equal to the size of a minimum clique partition of G'.

Finding a clique partition of a graph $G = \langle V, E \rangle$ is equivalent to finding a coloring of its complement $\overline{G} = \langle V, (V \times V) \setminus E \rangle$. This implies that the biclique cover number of a bipartite graph G corresponds to the chromatic number of $\overline{G'}$ [7].

5 A Concentration Result for Optimal Candidate Role-Sets

Using the model described in the previous section, we will prove that the cost of an optimal candidate role-set ψ for a given system configuration φ is tightly concentrated around its expected value. We will use the concept of martingales and the Azuma-Hoeffding inequality to obtain a concentration result for the chromatic number of a graph G [14,15]. Since finding the chromatic number is equivalent to both MCP and MBP, we can conclude that the minimum number of roles required to cover the user-permission relationships in a given configuration is tightly concentrated around its expected value.

Let G be an undirected unipartite graph, and $\chi(G)$ its chromatic number.

Theorem 2. *Given a graph G with n vertices, the following equation holds:*

$$\Pr(|\chi(G) - \mathbb{E}[\chi(G)]| \geq \lambda) \leq 2\exp\left(\frac{-2\lambda^2}{n}\right) \tag{3}$$

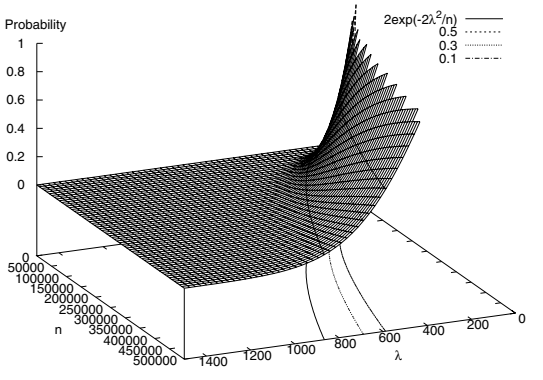
Proof. We fix an arbitrary numbering of the vertices from 1 to n. Let G_i be the subgraph of G induced by the set of vertices $1, \ldots, i$. Let $Z_0 = \mathbb{E}[\chi(G)]$ and $Z_i = \mathbb{E}[\chi(G) \mid G_1, \ldots, G_i]$. Since adding a new vertex to the graph requires no more than one new color, the gap between Z_i and Z_{i-1} is at most 1. This allows us to apply the Azuma-Hoeffding inequality, that is Equation 1 where $d_k = 1$.

Note that this result holds even without knowing $\mathbb{E}[\chi(G)]$. Informally, Theorem 2 states that the chromatic number of a graph G is sharply concentrated around its expected value. Since finding the chromatic number of a graph is equivalent to MCP, and MCP is equivalent to MBC, this result holds also for MBC. Translating these concepts in terms of RBAC entities, this means that the cost of an optimal candidate role-set of any configuration φ with $|UP| = n$ is sharply concentrated around its expected value according to Equation 3, where $\chi(G)$ is equal to the minimum number of required roles. It is important to note that n represents the number of vertices in the coloring problem but, according to the proposed model, it is also the number of edges in MBP; that is, the user-permission assignments of the system configuration.

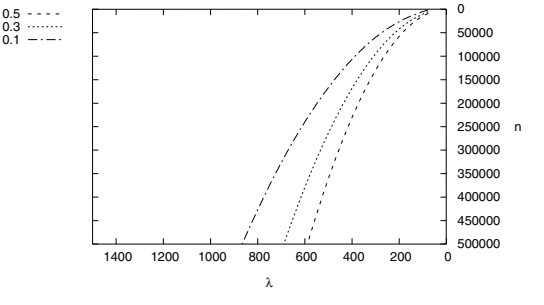
Figure 1(a) shows the plot of the Equation 3 for n varying between 1 and 500,000, and λ less than 1,500. It is possible to see that for $n = 500{,}000$ it is sufficient to choose $\lambda = 900$ to assure that $\Pr(|\chi(G) - \mathbb{E}[\chi(G)]| \geq \lambda) \leq 0.1$. In the same way, choosing $\lambda = 600$, then $\Pr(|\chi(G) - \mathbb{E}[\chi(G)]| \geq \lambda)$ is less than 0.5. Figure 1(b) shows the values for λ and n to have the left part of the inequality in Equation 3 to hold with probability less than 0.5, 0.3, and 0.1 respectively.

Setting $\lambda = \sqrt{n \log n}$, Equation 3 can be expressed as:

$$\Pr(|\chi(G) - \mathbb{E}[\chi(G)]| \geq \sqrt{n \log n}) \leq \frac{2}{n^2} \tag{4}$$



(a) Plot of Equation 3



(b) Highlight of some λ values for Figure 1(a)

Fig. 1. Relationship between the parameters λ, n and the resulting probability

That is, the probability that our approach differ from the optimum more than $\sqrt{n \log n}$ is less than $2/n^2$. This probability becomes quickly negligible as n increases. To support the viability of the result, note that in a large organization there are usually thousands user-permission assignments.

5.1 Applications of the Bound

Assuming that we can estimate an approximation $\tilde{\mathbb{E}}[\chi(G)]$ for $\mathbb{E}[\chi(G)]$ such that $|\tilde{\mathbb{E}}[\chi(G) - \mathbb{E}[\chi(G)]| \leq \varepsilon$ for any $\varepsilon > 0$, Theorem 2 can be used as a *stop condition* when striving to find an approximation of the optimum for any role mining algorithm. Indeed, suppose that we have a probabilistic algorithm that provides an approximation of $\chi(G)$, and suppose that its output is $\tilde{\chi}(G)$. Since we know $\tilde{\mathbb{E}}[\chi(G)]$, we can use this value to evaluate whether the output is acceptable and therefore decide to stop the iterations procedure. Indeed, we have that:

$$\Pr(|\chi(G) - \tilde{\mathbb{E}}(\chi(G))| \geq \lambda + \varepsilon) \;\leq\; 2\exp\left(\frac{-2\lambda^2}{n}\right).$$

This is because

$$\Pr(|\chi(G) - \tilde{\mathbb{E}}(\chi(G))| \geq \lambda + \varepsilon) \;\leq\; \Pr(|\chi(G) - \mathbb{E}(\chi(G))| \geq \lambda)$$

and, because of Theorem 2, this probability is less than or equal to $2\exp\left(-2\lambda^2/n\right)$. Thus, if $|\tilde{\chi}(G) - \tilde{\mathbb{E}}[\chi(G)]| \leq \lambda + \varepsilon$ holds, then we can stop the iteration, otherwise we have to reiterate the algorithm until it outputs an acceptable value.

For a direct application of this result, we can consider a system configuration with $|UP| = x$. If $\lambda = y$, the probability that $|\chi(G) - \mathbb{E}[\chi(G)]| \leq y$ is greater than $2\exp\left(-2y^2/x\right)$. We do not know $\mathbb{E}[\chi(G)]$, but since $|\tilde{\mathbb{E}}[\chi(G)] - \mathbb{E}[\chi(G)]| \leq \varepsilon$ we can conclude that $|\chi(G) - \tilde{\mathbb{E}}[\chi(G)]| < y + \varepsilon$ with probability at least $2\exp\left(-2y^2/x\right)$. For instance, we have considered the real case of a large size company, with 500,000 user-permissions assignments. With $\lambda = 1{,}200$ and $\varepsilon = 100$, the probability that $|\chi(G) - \tilde{\mathbb{E}}[\chi(G)]| < \lambda + \varepsilon$ is at least 99.36%. This means that, if $\tilde{\mathbb{E}}[\chi(G)] = 24{,}000$, with the above probability the optimum is between 22,700 and 25,300. If a probabilistic role mining algorithm outputs a value $\tilde{\chi}(G)$ that is estimated quite from this range, then it is appropriate to reiterate the process in order to find a better result. Conversely, let us assume that the algorithm outputs a value within the given range. We know that the identified solution differs, from the optimum, by at most $2(\lambda + \varepsilon)$, with probability at least 99.36%. Thus, one can assess whether it is appropriate to continue investing resources in the effort to find a better solution, or to simply accept the provided solution. This choice can depend on many factors, such as the computational cost of the algorithm, the economic cost due to a new analysis, and the error that we are prone to accept, to name a few.

There is also another possible application for this bound. Assume that a company is assessing whether to renew its RBAC state, just because it is several years old [19]. By means of the proposed bound, the company can establish whether it is the case to invest money and resources in this process. Indeed, if the cost of the RBAC state in use is between $\tilde{\mathbb{E}}[\chi(G)] - \lambda - \varepsilon$ and $\tilde{\mathbb{E}}[\chi(G)] + \lambda + \varepsilon$, the best option would be not to renew

it because the possible improvement is likely to be marginal. Moreover, changing the RBAC state requires a huge effort for the administrators, since they need to get used to the new configuration. In our proposal it is quite easy to assess if a renewal is needed. This indication can lead to important time and money saving.

Note that in our hypothesis, we assume that the value of $\tilde{\mathbb{E}}[\chi(G)]$ is known. Currently, not many researchers have addressed this specific issue in reference to a generic graph, whereas plenty of results have been provided for Random Graphs. In particular, it has been proven [13, 2] that for $G \in G_{n,p}$:

$$\mathbb{E}[\chi(G)] \sim \frac{n}{2\log_{\frac{1}{1-p}} n}$$

We are presently striving to apply a slight modification of the same probabilistic techniques used in this paper, to derive a similar bound for the class of graphs used in our model.

6 Conclusions and Future Works

In this paper we proved that the optimal administration cost for RBAC, when striving to minimize the number of roles, is sharply concentrated around its expected value. The result has been achieved by adopting a model reduction and advanced probabilistic tools. Further, we have shown how to apply this result to deal with practical issues in administering RBAC; that is, how it can be used as a stop condition in the quest for the optimum.

This paper also highlights a few research directions. First, a challenge that we are currently addressing is to derive an estimate of the expected optimal number of roles ($\mathbb{E}[\chi(G)]$) from a generic system configuration. Another research path is applying both the exposed reduction and the probabilistic tools to obtain similar bounds while simultaneously minimizing more parameters.

Acknowledgment

This work was partly supported by: The Spanish Ministry of Science and Education through projects TSI2007-65406-C03-01 “E-AEGIS” and CONSOLIDER CSD2007-00004 “ARES”, and by the Government of Catalonia under grant 2005 SGR 00446.

References

1. American National Standards Institute (ANSI) and InterNational Committee for Information Technology Standards (INCITS): ANSI/INCITS 359-2004, Information Technology – Role Based Access Control (2004)
2. Bollobás, B.: The chromatic number of random graphs. Combinatorica 8(1), 49–55 (1988)
3. Colantonio, A., Di Pietro, R., Ocello, A.: A cost-driven approach to role engineering. In: Proceedings of the 23rd ACM Symposium on Applied Computing, SAC 2008, Fortaleza, Ceará, Brazil, vol. 3, pp. 2129–2136 (2008)

4. Colantonio, A., Di Pietro, R., Ocello, A.: Leveraging lattices to improve role mining. In: Proceedings of the IFIP TC 11 23rd International Information Security Conference, SEC 2008. IFIP International Federation for Information Processing, vol. 278, pp. 333–347. Springer, Heidelberg (2008)
5. Coyne, E.J.: Role engineering. In: RBAC 1995: Proceedings of the first ACM Workshop on Role-based access control, Gaithersburg, Maryland, United States, p. 4. ACM, New York (1996)
6. Coyne, E.J., Davis, J.M.: Role Engineering for Enterprise Security Management. Artech House (2007)
7. Ene, A., Horne, W., Milosavljevic, N., Rao, P., Schreiber, R., Tarjan, R.E.: Fast exact and heuristic methods for role minimization problems. In: Proceedings of the 13th ACM Symposium on Access Control Models and Technologies, SACMAT 2008, pp. 1–10 (2008)
8. Frank, M., Basin, D., Buhmann, J.M.: A class of probabilistic models for role engineering. In: Proceedings of the 15th ACM Conference on Computer and Communications Security, CCS 2008, pp. 299–310 (2008)
9. Geerts, F., Goethals, B., Mielikäinen, T.: Tiling databases. In: Suzuki, E., Arikawa, S. (eds.) DS 2004. LNCS, vol. 3245, pp. 278–289. Springer, Heidelberg (2004)
10. Jajodia, S., Samarati, P., Subrahmanian, V.S.: A logical language for expressing authorizations. In: SP 1997: Proceedings of the 1997 IEEE Symposium on Security and Privacy, p. 31. IEEE Computer Society, Los Alamitos (1997)
11. Kuhlmann, M., Shohat, D., Schimpf, G.: Role mining – revealing business roles for security administration using data mining technology. In: Proceedings of the 8th ACM Symposium on Access Control Models and Technologies, SACMAT 2003, pp. 179–186 (2003)
12. Lu, H., Vaidya, J., Atluri, V.: Optimal boolean matrix decomposition: Application to role engineering. In: Proceedings of the 24th IEEE International Conferene on Data Engineering, ICDE 2008, pp. 297–306 (2008)
13. Łuczak, T.: The chromatic number of random graphs. Combinatorica 11(1), 45–54 (1991)
14. McDiarmid, C.J.H.: On the method of bounded differences. In: Siemons, J. (ed.) Surveys in Combinatorics: Invited Papers at the 12th British Combinatorial Conference. London Mathematical Society Lecture Notes Series, vol. 141, pp. 148–188. Cambridge University Press, Cambridge (1989)
15. Mitzenmacher, M., Upfal, E.: Probability and Computing: Randomized Algorithms and Probabilistic Analysis. Cambridge University Press, New York (2005)
16. Rymon, R.: Method and apparatus for role grouping by shared resource utilization, United States Patent Application 20030172161 (2003)
17. Schlegelmilch, J., Steffens, U.: Role mining with ORCA. In: Proceedings of the 10th ACM Symposium on Access Control Models and Technologies, SACMAT 2005, pp. 168–176 (2005)
18. Vaidya, J., Atluri, V., Guo, Q.: The role mining problem: finding a minimal descriptive set of roles. In: Proceedings of the 12th ACM Symposium on Access Control Models and Technologies, SACMAT 2007, pp. 175–184 (2007)
19. Vaidya, J., Atluri, V., Guo, Q., Adam, N.: Migrating to optimal RBAC with minimal perturbation. In: Proceedings of the 13th ACM Symposium on Access Control Models and Technologies, SACMAT 2008, pp. 11–20 (2008)
20. Vaidya, J., Atluri, V., Warner, J.: RoleMiner: mining roles using subset enumeration. In: Proceedings of the 13th ACM Conference on Computer and Communications Security, pp. 144–153 (2006)
21. Williams, D.: Probability with Martingales. Cambridge University Press, Cambridge (1991)
22. Zhang, D., Ramamohanarao, K., Ebringer, T.: Role engineering using graph optimisation. In: Proceedings of the 12th ACM Symposium on Access Control Models and Technologies, SACMAT 2007, pp. 139–144 (2007)