2015 Ninth International Conference on Complex, Intelligent, and Software Intensive Systems

Data Mining to Identify Fraud Suspected on Electronic Elections

Yuri Tadeu Poloni
dept. Information Security
University of Vale do Rio dos Sinos - UNISINOS
São Leopoldo, Brazil
e-mail: yuripoloni@gmail.com

Daniel Formolo
dept. Information Security
University of Vale do Rio dos Sinos - UNISINOS
São Leopoldo, Brazil
e-mail: danielformolo@unisinos.br

***Abstract*—Many democratic countries choose their representatives through electronic elections. Even being a modern tool, its results can be explored maliciously. Because that, many instruments and protocols are using to protect electronic elections from attacks. This work propose a new system to improve the security in electronic elections. It is based on analyses of behavior voter to detect electronic voting machines (EVM) with dissonant result, what can serve as start point to auditing. The system uses data mining to select suspect EVMs, presenting good results in the task of indicating EVMs with suspect of fraud.**

electronic voting machines; vote; behavior; auditing; data mining.

I. INTRODUCTION

There are many voting systems in the world, all of them aim to guaranty integrity and confidentially of the vote [1, 2, 3]. The necessity of secure election systems promotes discussions and new proposes to improve the electronic voting process [4, 5, 6, 7, 8].

Nowadays, there are three generations of voting systems, all of them are applied in different countries. Brazil uses the First Generation of electronic voting, where the EVM save the vote in its internal memory. An example of this model is the Direct Recording Electronic voting machine (DRE). Due to security reasons Paraguay, Germany and Netherland forbidden it [9].

The Second Generation improve the principle of software independence, saving the vote in memory and making a printed version of the vote for future check. Example of this kind of EVM is the Independent Voter-Verifiable Record (IVVR) and Voter-Verifiable Paper Audit Trail (WPAT), they are find in some US states [7].

Even with the evolution of electronic EVM to software independence, reported divergences between electronic results and printed votes continues an open problem. In case of a security fail, do not have evidences if the problem was in the software or printed votes [10]. This problem leads to Third Generation of electronic EVMs. That generation adds the possibility of voter verify his choice on the printed vote before it is putted in a backup EVM for future check [10]. Examples of this kind of model are End-to-End (E2E), Witness Systems (WS) and Electronic Vote Billet (BE). They are find in US and Argentina [10].

Even this system evolution is not considered secure enough for many countries, which prefer the old method, with manual vote/counting [5]. Due to identified security holes, researchers and security analyst suggest many protocols and methods to improve security in voting systems [3, 6, 8]. None method analyses the post-election vote, with objective of to identify fraud evidences in electronic elections. This work propose a new system based on data mining that identify possible frauds in electronic EVMs, facilitating the identification of fraud elections through suggestion of suspect EVMs, which can pass for post-election audit. The System just reports possible EVMs with fraud, the auditing and fraud proof are not in the scope of the work. The addition of this new security element to voting process difficults malicious actions, because forces to synchronize proportionally any action in all the EVMs of a region. That is explained in details ahead.

Next section explores in more details the relation of electronic EVMs, elections and security. The Section III brings an overview of pattern emerging from voters and how it can help to identify fraud suspects. The Section IV describes the proposed System. The research methodology is described in Section V, while Section VI shows the experiments and discusses their results. Finally, Section VII presents the conclusions.

II. ELECTION PROCESS AND SECURITY

DRE was the first electronic EVM model. It is very dependent of embedded software and of the counting process [11]. The security problem with that type of EVMs starts before the election. In general, the embedded code is not open, and none knows if the code has security drawbacks or malicious routines. Hours before start the election, a test is applied over some random selected EVMs. The goal is checks if the EVM memory is clean and the engine is running coherently, but there is no guaranties if someone updated the embedded software between the time interval of the test and the beginning of election. Adds to that security weakness the uncertainty about the embedded software in all other EVMs not tested, what are the big majority, besides other software and hardware drawbacks described in [12, 13].

After the election, the EVMs are moved to a defined local where data are transmitted to central counting system in security way. The moment of transmit the data is another opportunity to change them. Another insecurity spot is the counting process, there is not possibility to check if result is correct, because there is not a copy of the votes in another type of media [14].

978-1-4799-8870-9/15 $31.00 © 2015 IEEE
DOI 10.1109/CISIS.2015.8

IVVR and WPAT are the evolution of DRE, they print the votes, allowing each voter verify own choices. That is the first step to code independence[15]. Nevertheless, there is no guarantee against fails before and after the voting process. E2E, WS and BVE are different, including more security to the process. In case of BVE, for example, the ballot is divides in two parts; the attendant keeps with the first part, while the elector receive the second part, which has an RFID chip. The elector introduces his ballot into one of some available EVMs. At the end of voting process the data are recorded into the RFID chip and printed at the elector's ballot, who can compare the printed information with the recorded into the chip, just approaching the ballot of the EVM, which prints on the screen the recorded information of the chip. After that, the elector gives back his ballot to the attendant, who checks if the received ballot is the pair of his first part. Before end the voting process, the elector can request other ballots and restart his voting, in this case, old ballots are eliminated. At the end of entire voting process, the counting is started electronically using an EVM and the chip information of all second part of votes [7].

This process is more secure than other, none information is saved in the EVM, even the software can be checked at any moment, because it runs in a CD-ROM [10]. This voting process avoid many discussed drawbacks [11]. Nevertheless, in practice, most of countries uses other systems, and even DRE has some weakness.

All described electronic voting systems and any other need cover some essential principles. They are: Confidentiality, where voter identity must be preserved; Availability, that is the guarantee of access and use of the system by the authorized people; and Integrity, where the original data must be preserved [16]. Therefore, all voting systems try to cover these principles to avoid fails. Most of the efforts focus in the voting process, some examples can be seen in [6, 8, 17], but no one uses the voting results to point out fraud suspects. Data mining of voting results are used to analyze elector characteristics [18, 19]. A similar approach of the System is found in Cabral`s work [19]. It tracks, with success, energy consume habit to detect fraud in electrical systems. The patterns found in voting behavior can be expanded to online voting systems. The voters manifest their opinion through a cloud system, but they are also part of a community in a physical place. In that way, they practice and suffer influence of people who live at same region. The big problem for online voting systems is identify from which region each voter belongs. Even identifying the voter and his region by a form, or localize the region where the vote was sent, one can hack the vote region and baffle an anti-fraud system. Off course, the systems become hard the hacker`s life, because there is one more level of protection to it takes care.

III. VOTING BEHAIVOR AND DATA CHARACTERISTICS

Individuals who belongs to the same familiar, religious, professional, friendship or neighborhood tends to think likewise [21]. Consequently, their ideas about policy are similar [13, 22, 23], local economic interesting enforces that characteristic. Also according to Levernier [24] regional location, economy and population density are important characteristics in voting decision. That pattern are show in India voting patterns [13].

It is natural expects similar voting behavior for close localities. Therefore, an way to identify possible frauds is looking for EVM sites whose vote result differs from other close EVM sites. In general, electronic voting process gives percentage of each candidate voting by EVM and the exact EVM location. Based on the above idea, this data are the minimum to identify EVMs out of expected pattern. The simple cross of this data is not enough because the emerged pattern has variations explained ahead and the voting percentage of each candidate oscillates more or less according to the importance of each candidate in the slice of electors. For example, few representative candidates have low number of votes and two EVMs located at same site can present big percentage divergence for the same candidate even with few difference of absolute number of votes between the EVMs. By the other hand, voting percentage of weigh candidates are no influenced by few vote differences between close EVMs.

Considering the data characteristics and the objective of work, a good approach is explores the voting data with a clustering data mining algorithm. That algorithms sort out data in sets according to defined characteristics [25]. They can be divided in seven groups [26]: Hierarchical Clustering, Partitioning Relocation Clustering, Density-Based Partitioning, Grid-Based Methods, Co-Occurrence of Categorical Data and Other Clustering Techniques. We opt by use a Density-Based Partitioning algorithm called DBSCAN, because the problem consist of grouping EVMs distributed on a continuum Euclidian space, based on their similar characteristics [26]. It is expected find only one single set to each candidate in each analyzed region. More than one set indicates a possible fraud, due to the characteristics of data set that, in general, presents similar voting percentages for close EVMs.

The algorithm DBSCAN is a traditional density algorithm. It is good to identify noise elements and identify elements with similar characteristics. Even if this characteristics change along of Euclidian space. For this type of data, there are two main observed behaviors, first is shown in Figure 1, where one or more candidates have much influence in the same region. In this situation, two centers are formed and intermediate region is mixed. The example shows Candidate A and B, each one with their influence region defined and region C with mixed influence. Other common situation is a big region influenced by one candidate, Figure 2 show that situation. The candidate earn many votes in center of influence, as far as we move away from center of influence, his votes percentage reduces.

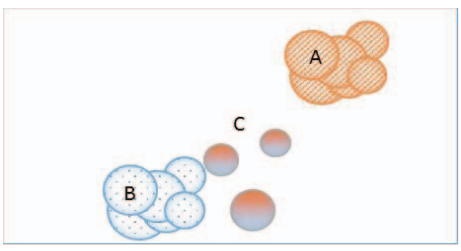

Figure 1. Influence example of two strong candidates in close regions.

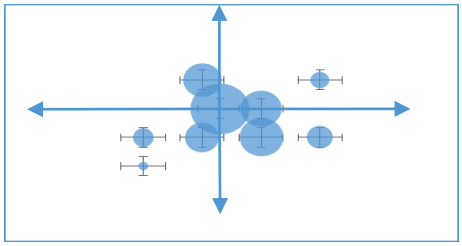

Figure 2. Influence example of one strong candidates in many close regions.

These types of patterns are well identified by DBSCAN, in case of Figure 2, it identifies one single set, even with big vote difference between central and peripheral EVMs, because DBSCAN understand the pattern of natural reduction of votes along of Euclidian space. Any other identified set will be a suspect group of EVMs that do not follow the regional pattern. The example in Figure 1 is more complex than shown in Figure 2, two sets are identified by DBSCAN and the group of EVMs C will be divided between sets A and B. Eventually some EVMs will be classified as noise, do not appearing in any set or sometimes a new set C is created.

IV. PROPOSED SYSTEM

In Brazil are around 312 thousand EVMs and 140 million electors [27], so in Brazil and generally other countries just some EVMs are selected to be auditing [28]. The impossibility of auditing all EVMs and the weakness of electronic voting systems, especially the first generation motived the creation of this new system, which helps select EVMs for auditing.

Figure 3 shows the modules of the System[1]. Module A filters in database the city to inspected, all EVMs location and vote percentage of all candidates for such selected position. Module B controls the process, starting the process for each candidate of selected election. The next stage applies DBSCAN algorithm to analyses the vote percentage behavior of the candidate indicates by Module B in all EVMs of the city. The Module D compiles the pre-results and the passes the control to Module B again. If there are more candidates to be analyzed, it selects the next candidate, his vote data and restarts Module C. Module D adds new results to the report and when no more analyses are required, the control passes to Module E that shows the result. The report shows all possible EVMs with fraud classified by candidate.

The Module D decides what EVMs are suspect. In a neighborhood, only one group of EVMs must be identified. More than one group could mean that one or more groups are fraud because the expected pattern is not followed. The other possibility is just a natural result that does not follow the pattern. The System provides an indication of fraud, but a forensic analysis must be made after that.

V. METHODOLOGY

The data used to validate the system came from Brazilian election 2010 and are available by TSE[2]. This work considers only the results of president position. The attributes of EVMs are normalized by the system, that is necessary to the system avoid the magnitude of distortions [25].

To work, DBSCAN requires two parameters: φ refers to the minimal number of elements needed to create a set, while ε represents the maximum distance among the elements that compose a set [25]. The System uses a library [3] that implements DBSCAN algorithm, its version is 2.0. The voting percentage of a candidate defines the elements in a DBSCAN set, so the distance between closest element of the last element in a set must be less than ε.

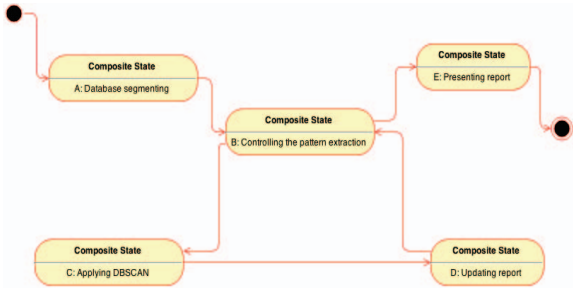

Figure 3. UML State Diagram of the System.

In other cases, the elements compose other sets or a noise set. For this work was defined ε=6. To find this value we run the System to all candidates, for each test varying the ε from 2 to 10, comparing the number of noise sets generated. The result is shown in Figure 4. When ε raises, we observe the reduction of noise sets, but we know that big ε means join of sets and consequently the reduction of sensibility. Therefore, the best is keep ε low. By the other hand, low ε means many noise sets and many elements classified in noise set. We can observe that the decrease of noise sets is exponential so, for this application, when ε>6, the reduction of noise sets is low, what means that a ε=6 is a good tradeoff.

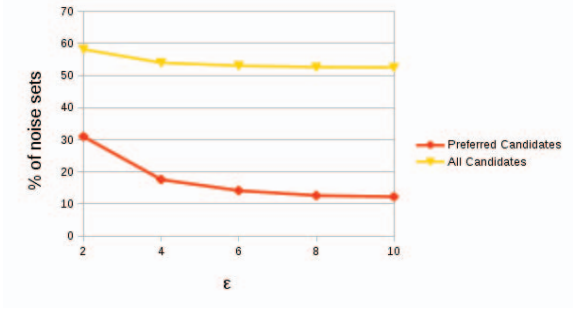

Figure 4. Variation of noise sets generated by DBSCAN, according to variation of ε.

[1] The system: http://projetofinal.polonilabs.com.br:3000/ Source code: https://bitbucket.org/yuripoloni/projeto_final

[2] Tribunal Superior Eleitoral. All election data are available in: http://www.tse.jus.br/eleicoes/eleicoes-anteriores/eleicoes-2010/estatisticas

[3] DBSCAN library is available in: https://github.com/matiasinsaurralde/dbscan

VI. EXPERIMENTS AND RESULTS

The first experiment analyses number of group patterns generated in a city with more than 220 thousand habitants. The Figure 5 presents the sets generated by the System from some neighborhoods of Caxias do Sul city. The neighborhoods are divided in 3 groups: Far downtown, Close Downtown and Downtown.

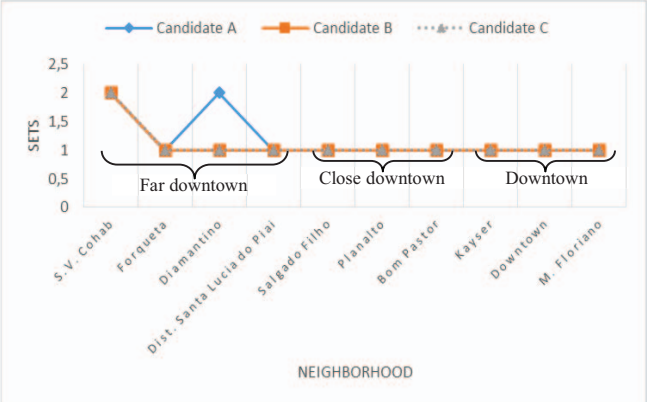

Figure 5. Sets extracted by neighborhood from the three main presidential candidates.

The sets represent the found patterns for the three main presidential candidates, for all of the candidates there is the expected pattern of 1 set by neighborhood. For the 3 candidates, some exceptions occurred in neighborhoods far from downtown where up to 2 sets were found. That means 2 groups with different patterns of voting percentage for each candidate. In S.V. Cohab and Diamantino, the System reported possible frauds, what do not mean a real fraud but some strange behavior. For the low-ranked candidates, with less than 120 thousand voters each in whole country, the system do not found patterns and become impossible finds fraud evidences.

The same test was applied to other 9 cities with similar results, including finding a little bit more sets generated from far downtown.

As observed in Figure 5, in general, the system finds a pattern in neighborhoods for principal candidates and eventually some EVMs are separated in 2 or more sets. Preferred candidates are easier to track than other candidates. That is clear when observing ε in Figure 4, the number of noise sets to preferred candidates are less than all candidates.

The Figure 6 is part of report related to data shown in Figure 5. It shows the analyses of one of the main candidate in Diamantino neighborhood. Many EVMs are not shown by the System because they compose a single set, i.e., all of them have the same voting pattern. However, the system tracked 2 patterns of reported EVMs. One pattern ranges from 38% to 45% of votes. The other is composed by an isolated EVM classified in noise set (Grupo -1), with a very different pattern of 30% of votes. What suggest firstly a real different pattern or set -1 defrauded or even set 0 defrauded. Based in this data, a person can be more assertive to select EVMs to audit.

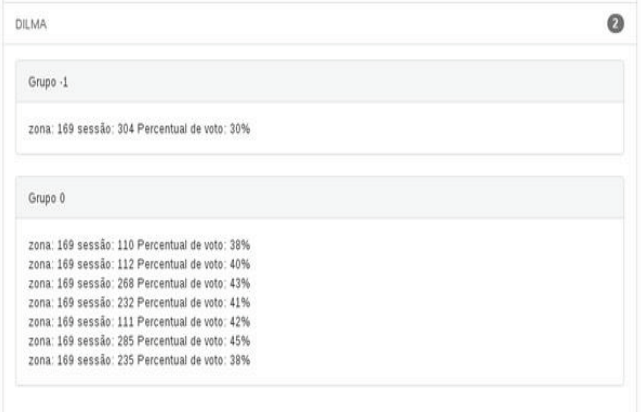

Figure 6. Report segment of fraud suspect.

Due the difficult to finds election database with fraud EVMs, the second block of experiments uses a biased database with well-defined situations showed in TABLE I. The system responds according expected for all situations. For situations A the EVMs clearly form a single set, therefore, making just one set and there are not fraud evidences. For situation B even with a large distance between the EVM with 13% of votes and the EVM with 6% of votes, there is a pattern of loss votes along EVMs are located away. That reflects the situation depicted in. For situation C, there are 2 groups of EVMs and the system reports these two sets as fraud suspect. Situation D is more complex than the other. After many experiments varying randomly the vote percentage, the System always identify a fraud suspect when the distance of vote percentage among two or more EVMs exceeds 6%, that is exactly the ε configured in DBSCAN. That expected result shows a limitation of the System. Swindling EVMs, with vote variation less than 6%, are not identified.

TABLE I. Description of data evaluated by the system set with Φ=1 and ε =6.

Situation	Description	Fraud Suspect
A	Eight EVMs varying randomly the vote percentage of observed candidate from 59% to 61%	No.
B	Eight EVMs decreasing the vote percentage of observed candidate sequentially from 13% to 6%	No.
C	Eight EVMs divided in 2 groups of vote percentage of observed candidate. First: varying randomly from 39% to 40%. Second: varying randomly from 22% to 23%.	Yes: two sets identified.
D	Eight EVMs varying randomly the vote percentage of observed candidate from 10% to 20%	Yes: one or two sets, depending on percentage distance among EVMs result.

VII. CONCLUSION

Fraud in electronic elections are a constant risk, for that reason, there are many proposes solutions to protect voting systems. The majority of them focused to avoid attacks and data manipulations. That is very important, but no one focused in indicate fraud suspects, according to election results. This work presented a System based on data mining, which analyses the results of electronic elections, identifying fraud suspect EVMs. The use of the System can help auditing process to select the suspect EVMs and voting sites with fraud suspect. More than electronic EVMs, there is the possibility of expand the system for online voting systems, and tracks possible hacking votes that are out of regional vote patterns.

The System shows promising results, although has limitations to points fraud suspects that change the original results in less than 6% for a candidate, and specially to analyze performance of candidates with low number of votes.

It is necessary more experiments with other database elections to validate its performance and potentials. Besides evaluate different election databases, it is important explore other clustering algorithms to solve the problem. To conclude, the System can expand his use potential if applied to analyzes electors behavior aspects of the different voting regions.

REFERENCES

[1] P. N. Neumann, "Security criteria for electronic voting," 16th National Computer Security Conference, September 1993.

[2] A. Riera, P. Brown, "Bringing confidence to electronic voting" Electronic Journal of e-Government vol. 1, Issue 1, pp. 14–21, 2003.

[3] B. Randell, P. Ryan, "Voting technologies and trust", Formal Aspects in Security and Trust, vol. 3866 pp. 1–4, 2006.

[4] A. D. Rubin, "Security considerations for remote electronic voting", Commun. ACM, vol. 45, pp. 39–44, December, 2002.

[5] D. De Cock, P. Bart, "Electronic voting in Belgium: past and future", E-Voting and Identity, vol. 4896, pp. 76–87, 2007.

[6] S. M. Abdulhamid, S. A. Olawale, "The design and development of real-time EVoting", Computer Network and Information Security, pp. 9–18, April 2013.

[7] B. Goldsmith, "Electronic voting & counting technologies: a guide to conducting feasibility studies", International Foundation for Electoral Systems, June 2011.

[8] A. F. N. Al-Shammari, A. Villafiorita, "A synthesis of vote verification methods in electronic voting systems", Design, Development, and Use of Secure Electronic Voting Systems, 2014.

[9] A. F. Brunazo, "Fraud and defenses in the electronic voting - Fraudes e defesas no voto eletrônico", Ed. All Print, 2006.

[10] CMIND, "Election observation report in Argentina - Relatório da observação de eleição na Argentina", October 2011.

[11] C. Enguehard, "Transparency in electronic voting: the great challenge, IPSA International Political Science Association RC 10 on Electronic Democracy. Conference on "E-democracy - State of the art and future agenda, January 2008.

[12] R. Gonggrijp, W. J. Hengeveld, "Studying the Nedap/Groenendaal ES3B voting computer: a computer security perspective", Proceedings of the USENIX Workshop on Accurate Electronic Voting Technology, 2007.

[13] A. Feldman, J. A. Halderman, E. W. Felten, "Security analysis of the Diebold AccuVote-TS voting machine", Electronic Voting Technology Workshop (EVT'07), 2007.

[14] E. A. Fisher, K. J. Coleman, "The direct recording electronic voting machine (DRE) controversy: FAQs and misperceptions", Congressional Research Service, December 2005.

[15] R. L. Rivest, J. P. Wack, "On the notion of software independence in voting systems", 2006.

[16] A. F. Brunazo, "Burla Eletrônica – The machine that makes your vote disappear - Burla Eletrônica – A máquina que faz seu voto sumir", Editora Fundação Alberto Pasqualini, 2002.

[17] A. B. Sangar, S. R. Khaze, L. Ebrahimi, "Participation anticipating in elections using data mining methods", International Journal on Cybernetics & Informatics (IJCI) vol. 2, n. 2, April 2013.

[18] B. Schneier, "Applied cryptography: protocols, algorithms, and source code in C", Ed. John Wiley & Sons, Inc., 1993.

[19] T. M. Kodinariya, R. Seta, "Visual data mining in indian election system", International Journal on Computer Science and Engineering, vol. 4, issue 7, 2012.

[20] J. E. Cabral, M. E. Gontijo "Fraud Detection in Electrical Energy Consumers Using Rough Sets", Systems, Man and Cybernetics, 2004 IEEE International Conference on, vol. 4, pp. 3625-3629, 2004.

[21] M. M. Castro, "Determinants of Electoral Behavior: The Centrality of Sophistication Policy - Determinantes do Comportamento Eleitoral: A Centralidade da Sofisticação Política" PhD tesis , 1994.

[22] C. Zucco, "Poor voters vs. poor places: persisting patterns in presidential elections in Brazil", Metropolis and Inequalities Seminar, Sao Paulo, 2010",

[23] H. Kitschelt, S. Wilkinson, "Patrons, clients, and policies: patterns of democratic accountability and political competition", Cambridge University Press, 2007.

[24] W. Levernier, A. G. Barilla, "The effect of region, demographics, and economic characteristics on county-level voting patterns in the 2000 presidential election", The review of regional studies, vol. 36, n. 3, pp. 427–447, 2006.

[25] P. Tan, M. Steinbach, V. Kumar. Introduction to Datamining, May 2005.

[26] P. Berkhin, "A survey of clustering data mining techniques", Grouping Multidimensional Data, pp. 25-71, 2006.

[27] TSE, "Electoral Brasilian Code, Law n. 4.737, of July, 15 of 1965", acessado em 13 de março de 2014.

[28] TSE, "Inside the e-voting machine - Por dentro da urna" ed. 2, revision e, 2010.