

Target Vehicle Identification for Border Safety using Mutual Information

Siddharth Kaza, *Member, IEEE*, Tao Wang, Hemanth Gowda, and Hsinchun Chen, *Senior Member, IEEE*

Abstract— The security of border and transportation systems is a critical component of the national strategy for homeland security. The security concerns at the border are not independent of law enforcement in border-area jurisdictions because information known by local law enforcement agencies may provide valuable leads useful for securing the border and transportation infrastructure. The combined analysis of law enforcement information and data generated by vehicle license plate readers at the international borders can be used to identify suspicious vehicles at ports of entry. This not only generates better quality leads for border protection agents but may also serve to reduce wait times for commerce, vehicles, and people as they cross the border. In this paper we use the mutual information concept to identify vehicles that frequently cross the border with vehicles that are involved in criminal activity. We find that the mutual information measure can be used to identify vehicles that can be potentially targeted at the border.

I. INTRODUCTION

HOMELAND security concerns in recent years have identified border and transportation security as critical areas. The national strategy for homeland security [1] calls for the creation of “smart borders” that provide “greater security through better intelligence, coordinated national efforts, and unprecedented international cooperation against the threats posed by terrorists, the implements of terrorism, international organized crime, illegal drugs, illegal migrants, cyber crime, and the destruction of natural resources.” In addition, the report also emphasizes that information sharing

Manuscript received March 7, 2005. This research was supported in part by the NSF Digital Government (DG) program: “COPLINK Center: Information and Knowledge Management for Law Enforcement” #9983304, NSF Knowledge Discovery and Dissemination (KDD) program: “COPLINK Border Safe Research and Testbed” #9983304, NSF Information Technology Research (ITR) program: “COPLINK Center for Intelligence and Security Informatics Research - A Crime Data Mining Approach to Developing Border Safe Research” #0326348, and Department of Homeland Security (DHS) and Corporation for National Research Initiatives (CNRI) through the “BorderSafe” initiative #2030002.

S. Kaza is with the Department of Management Information Systems at the University of Arizona, Tucson, AZ 85721 USA (phone: 520-621-2165; fax: 520-621-2433; email: skaza@eller.arizona.edu).

T. Wang is with the Department of Management Information Systems at the University of Arizona, Tucson, AZ 85721 USA (email: twang@eller.arizona.edu).

H. Gowda is with the Department of Computer Science at the University of Arizona, Tucson, AZ 85721 USA (email: heman@email.arizona.edu).

H. Chen is with the Department of Management Information Systems at the University of Arizona, Tucson, AZ 85721 USA (email: hchen@eller.arizona.edu).

systems are the foundations to improve the nation’s infrastructure.

The Department of Homeland Security (DHS) monitors vehicles entering the country, recording their license plates with a date and time of entry using automatic license plate readers. Customs and Border Protection (CBP) agents search vehicles for drugs and other contraband. These thorough checks are done for vehicles on watch lists (target vehicles) and on random vehicles as well. This process is time consuming and if the waiting times become too long, the flow of commerce, vehicles and people is impaired.

Better quality target vehicles can be identified at the border by utilizing information from other law enforcement agencies. The criminal associations of vehicles crossing the border are recorded in local law enforcement databases in border-area jurisdictions [2]. Customs and Border Protection agents believe that vehicles involved in illegal activity (especially smuggling narcotics) operate in groups. Thus, if the criminal links of one vehicle in the group are known, then their border crossing patterns can be used to identify other partner vehicles. In addition, vehicles that may not have any link to criminal activity may often cross together too (people going to work at the same time every morning). Only if one of the vehicles is linked to criminal activity then the other partner vehicles become interesting. Thus, law enforcement data can be used as a good anchor to perform association analysis.

We propose to perform this association analysis by using the concept of mutual information to identify pairs of vehicles that are potentially involved in criminal activity together. We ask the question: can law enforcement information from border area jurisdictions be used to identify target vehicles at the border? To test the implementation of a mutual information based algorithm we use local law enforcement data from southern Arizona along with border crossing data at the Arizona ports of entry. We find that the mutual information measure can be used to identify vehicles that can be potentially targeted at the border.

In the next section, we discuss background information and previous studies using mutual information. Section III explains the research testbed and research design. Experimental results are shown in Section IV and discussed in Section V. Section VI concludes and presents future directions.

II. LITERATURE REVIEW

A. Integration of Data from Multiple Sources

In order to explore the criminal links of border-crossing vehicles it is necessary to extract data from multiple sources. To triangulate information about a vehicle, it is necessary to reconcile all the instances of the vehicle across datasets which is a challenging task. Matching of entities and their relationships is a task that is hampered by problems that include [3]: *name differences*: similar entities in different databases have different names, *missing and conflicting data*: incomplete data or different values in different sources, and *object identification*: lack of global identifiers.

We use the BorderSafe information sharing and analysis framework [2] for accessing information from multiple datasets. The vehicles in multiple law enforcement datasets are matched by license plates and their criminal incidents are extracted.

B. Mutual Information

The concept of mutual information has been used to identify interesting co-occurrences of objects in many different application domains. One of the earliest definitions of mutual information was given by Fano [4] who defined it as the amount of information provided by the occurrence of an event (y) about the occurrence of another event (x). The mathematical equation is given by [4]

$$I(x; y) = \log \frac{P(x, y)}{P(x)P(y)}$$

Intuitively, this concept measures if the co-occurrence of x and y is more likely than their separate occurrences. The measure is positive if x and y are more likely to occur together ($P(x, y)$) than separately ($P(x)P(y)$). The concept (that Fano [4] defined for binary digits) was ported to studying mutual information between words in English texts [5],[6] and phrase extraction in Chinese [7]. Mutual information works well in text as each document can be considered as a set of events (words) and the probability of an occurrence of a word can be calculated over the entire document.

Mutual information based techniques have also been used for extracting protein motif patterns from protein sequences [8] and for identifying fundamental building blocks of proteins [9] from biomedical abstracts.

Since border-crossing records can be considered as a stream of text (license plates) implicitly ordered by the time of crossing, the mutual information measure can be used to identify co-occurrence between two crossing records. We propose the use of this measure to identify vehicles that cross together frequently.

III. RESEARCH TESTBED AND DESIGN

A. Testbed

The testbed for this research includes datasets obtained from Tucson Police Department (TPD), Pima County Sheriff's Department (PCSD), and Customs and Border Protection (CBP). The TPD and PCSD datasets contain police incidents from 1990 to 2005. These incidents include individuals and vehicles involved in illegal activity in most of southern Arizona. A summary of these datasets is shown in Table I.

TABLE I
KEY STATISTICS OF TPD AND PCSD DATA

	TPD	PCSD
Recorded Incidents	3.3 million	2.18 million
Vehicles	800,656	520,539

CBP data includes information on vehicles crossing the border between Arizona and Mexico at six ports of entry. This data includes the license plate, state, date, and time for crossings from 2003 to 2004. In addition, CBP also provides us with a list of vehicles seized at the border for illegal activity in 2003. Details of these datasets are shown in Table II.

TABLE II
SUMMARY OF BORDER CROSSING AND SEIZED VEHICLES

Recorded crossings	7.5 million
Number of Vehicles	1.2 million
Number of Seized Vehicles	530

B. Research Design

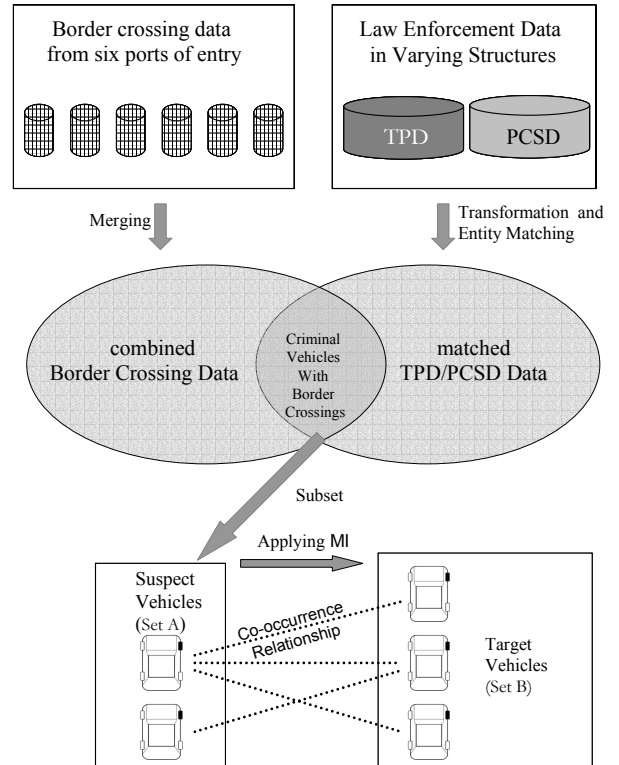


Fig. 1. The steps in the process to identify target vehicles

Fig. 1 shows the steps in the process of utilizing

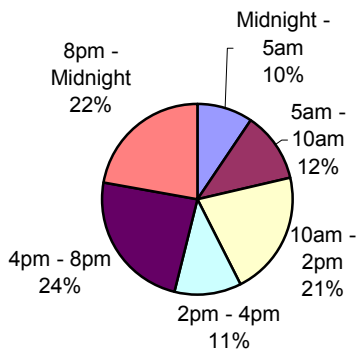


Fig. 3a. Distribution of crossings over times in a day

information from multiple jurisdictions to identify target vehicles at the border. First, TPD and PCSD data in varying database structures is transformed into a common schema. The vehicles in both these jurisdictions are matched to produce a combined vehicle list. More details on this procedure have been discussed in [2]. The border crossing information from all six ports of entry is consolidated. The vehicles in the TPD/PCSD data are then matched by license plates and states to the border crossing records to identify the crossings associated with them.

Using the process in Fig.1 we identified a subset of vehicles that were arrested (with individuals) for illegal activity in the jurisdictions of TPD/PCSD since January 2000. Since we are interested in associates of frequent border crossers, we further restricted the set to contain only the vehicles that have crossed the border more than 10 times. Hereafter, this subset will be referred to as ‘Set A’. Using the mutual information measure described in the next subsection, we identify other vehicles that cross the border frequently with the vehicles in Set A.

To utilize information from local jurisdictions to enhance the value of border crossing information we need to ascertain the amount of overlap that exists between the datasets. We did some exhaustive overlap analysis and some preliminary analysis on the temporal changes to the number of border crossings, which is presented in the next section. Such analysis facilitates the identification of high intensity border crossing periods. This information can be used to inform mutual information algorithms to assign higher and lower weights to certain associations.

C. Mutual Information Score

The mutual information score between any two vehicles is measured by using the formula:

$$MIS(v_1, v_2) = \log_2 \frac{P(v_1, v_2)}{P(v_1)P(v_2)}$$

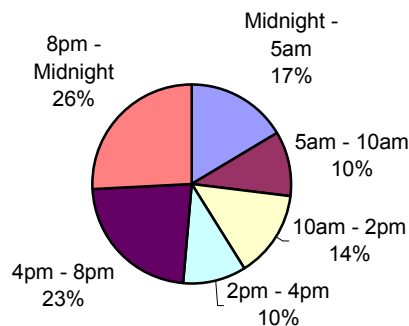


Fig. 3b. Distribution of the time of crossing of interesting pairs.

Here v_1 is a vehicle in Set A and v_2 is a vehicle that crosses within one hour of v_1 . $P(v_1, v_2)$ is the probability of v_2 crossing within one hour of v_1 , this is calculated based on the number of times v_1 and v_2 are seen crossing together. $P(v_1)$ is the probability of the vehicle v_1 crossing the border, which is defined by number of crossings of v_1 divided by total number of crossing records. Similarly, $P(v_2)$ is the probability of vehicle v_2 crossing the border. The \log_2 function is used to reduce the score to a small range.

IV. EXPERIMENTAL RESULTS

A. Overlap Analysis and Basic Statistics

The overlap between the border crossing data and local law enforcement data is shown in Table III. The values in the cells are the number of vehicles common to the datasets. The numbers suggest that many vehicles crossing the border have criminal incidents recorded in local law enforcement databases. This is a positive sign since it allows us to identify target vehicles at the border by exploring their criminal links.

TABLE III
OVERLAP BETWEEN TPD, PCSD, BORDER CROSSINGS, AND SEIZED VEHICLES

	TPD / PCSD
Border Crossing Vehicles	36,400
Seized Vehicles	65

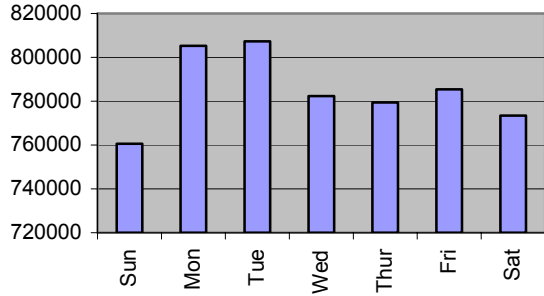


Fig. 2. Number of border crossings over different days of the week.

Fig. 2 shows the distribution of border crossings over different days of the week. It can be seen that most of the border crossings happen on weekdays, which can be attributed to work related traffic. Fig.3a presents the distribution of crossings by six different time slots in a day. These time slots were selected to reflect patterns of activity that were suggested by CBP agents. The morning 5 am – 10 am slot contains people usually going to work in Mexico or coming for work in the U.S. Similarly, the 4 pm – 8 pm slot usually signifies people returning from work. The 10 am – 2 pm slot points to people going for lunch. It can be seen that most of the traffic is during the working hours. We will refer to the Fig. 3b later in the paper.

B. Results based on Mutual Information Scores

For each vehicle in *Set A*, its mutual information (MI) scores with all the vehicles that crossed within one hour of *Set B* were calculated. This generated scores in the range of 11 to 22 for 467,000 pairs of vehicles. Since MI scores are calculated for each pair of vehicles, we need to identify a score threshold to extract interesting pairs. Previous research using mutual information [5],[7] have used experiments or manual inspection to establish such thresholds. We made a preliminary attempt to identify the threshold by gauging the overlap between vehicles in *Set B* with vehicles that were known to be criminal (vehicles in TPD/PCSD and vehicles seized by CBP). To determine this threshold we calculated the ratio

$$\frac{\text{Number of vehicles with criminal records}}{\text{Total vehicles}}$$

for all the vehicle pairs in *Set B* with MI scores greater than the threshold. Maximizing this ratio allowed us to select a threshold that resulted in the most vehicles with criminal records with the least amount of noise. We found that setting the value of the threshold to 16 identified 2,100 vehicles with criminal incidents in TPD/PCSD. Threshold values above and below 16 caused a sharp drop in the ratio. Even though this is a rough approximation, it allowed us to do some interesting analysis.

Table IV shows the overlap of vehicles above the threshold in *Set B* with TPD/PCSD and seized vehicle information.

TABLE IV
OVERLAP OF VEHICLES IN SET B (IDENTIFIED USING MI SCORES)

TPD / PCSD	Seized Vehicles
2,100 Incidents	33
160 Narcotics Crimes	

Since the 2,100 vehicles identified cross often with criminal vehicles (in *Set A*) and possess criminal records themselves, they make potential targets that should be investigated at the border. In addition, the overlap with seized vehicles is high which suggests that the vehicles identified using this method have strong criminal associations. However, in addition to maximizing the number of vehicles with criminal records, we also need to minimize the noise (i.e., identify less number of vehicles that might not be involved in criminal activity). A discussion of this aspect is presented in Section V and committed to future work.

Since, the vehicles identified by the MI method are promising with respect to their links to crime; further investigation into their patterns of crossing the border is warranted. Analysis of the patterns can help us improve the mutual information technique and apply it better to this domain. For instance, if most of the vehicles cross in a certain time period, then we can extend the MI measure to assign more weight to crossings in that period. With this aim in mind we analyzed the characteristics of the interesting pairs that were above the threshold value. Fig. 3b shows the distribution of the times when the pairs of vehicles in *Set A* and *Set B* crossed.

On comparing Fig. 3a to Fig. 3b, it can be seen that though only 32% of all crossings are recorded between 8 pm and 5 am, the interesting pairs of vehicles cross 43% of the time in the same period. If the time period is extended to 7 pm to 5 am, the number jumps up to 50% (as compared to 38% for all vehicles). Even though more analysis is needed, these numbers suggest that a large number of the pairs of vehicles with criminal activity cross the border during the night. This fact can be used to inform mutual information based algorithms.

Fig. 4 presents the distribution over weekdays; the distribution does not show any surprising differences with

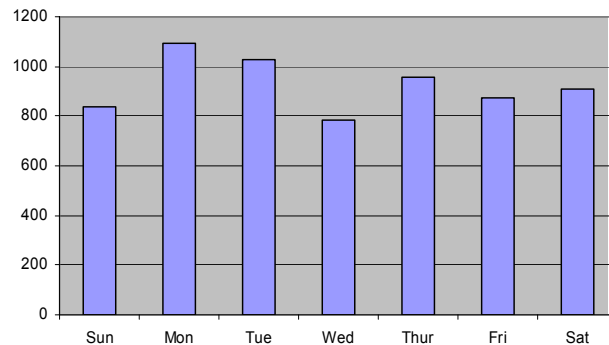


Fig. 4. Distribution over weekdays of interesting pairs in Set B

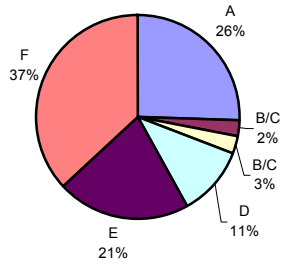


Fig. 5a. Distribution over ports (A-F) of all border crossing vehicles.

respect to the distributions of all border crossings (Fig. 2). Fig. 5a and Fig. 5b show the distribution of ports of entry for all border crossings and the interesting pairs respectively. Here again, it can be seen that the criminal pairs of vehicles seem to prefer crossing at Port E (security considerations prevent us from releasing real port names) more than other ports. This phenomenon may be due to the proximity of that port to certain areas in Mexico or US, or due to other reasons. The information can be used to enhance the mutual information algorithm by assigning more weight to frequent crossings at Port E.

C. Some Striking Examples

We analyzed the relationships between some of the pairs of vehicles with a high mutual information score to verify if they were interesting associations. An example of one such relationship was found between a pair of vehicles that had a mutual information score of 19.17. The first vehicle (hereafter Vehicle A) has 3 narcotics incidents in TPD/PCSD. The associated vehicle (hereafter Vehicle B) identified by the algorithm was found to have 2 narcotics incidents. Fig. 6 presents a distribution of the crossing times of the two vehicles. It can be seen that the vehicles crossed within 1 hour of each other a total of 12 times which might point to a partnership between them. Thus given the knowledge of the criminal incidents of Vehicle A, the mutual information algorithm could identify a potential

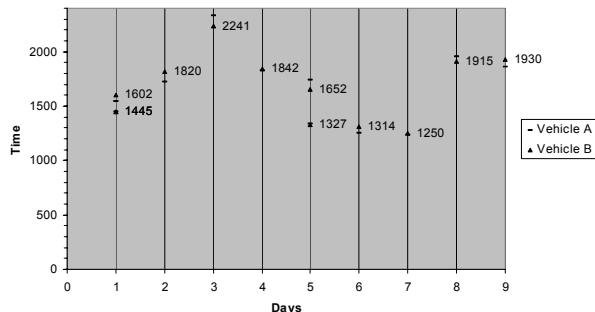


Fig. 6. Time distribution of crossings of Vehicle A and Vehicle B

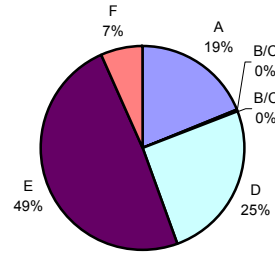


Fig. 5b. Distribution over ports (A-F) of interesting pairs in Set B.

target vehicle (Vehicle B) at the border. Even though the algorithm needs to be tested thoroughly, this and other examples point to the usefulness of the mutual information measure in this domain.

V. DISCUSSION

A threshold value of 16 for the MI score identified 2,100 vehicles with past criminal records. However; it also ascertained many other vehicles that had no past criminal records. Even though this might not look promising in other domains, it has positive connotations in this one. It suggests that many of the vehicles postulated to be criminal by the algorithm were not known to have criminal links before. So the algorithm can be used to identify potentially new criminal vehicles that can be targeted at the border. The low number of criminal vehicles might also be a result of the properties of the datasets. Most of the border crossing vehicles in our datasets may be headed for Phoenix, AZ and surrounding areas and thus their criminal activity will be recorded in those datasets. A more accurate test of the algorithm is possible if those datasets were available.

Overall, the mutual information measure is promising for this scenario but needs to be extended to include domain and contextual information (like time of crossing) to be effective.

VI. CONCLUSIONS AND FUTURE DIRECTIONS

Exploring the criminal links of border crossing vehicles in local law enforcement databases can be used to enhance border and transportation security. In this study we explored the characteristics of border crossing activity to better understand their temporal patterns. We also used the mutual information measure to identify vehicles that frequently cross with vehicles having criminal associations. We found that the mutual information measure can be used to identify vehicles that can be potentially targeted at the border.

In the future we intend to extend the mutual information measure to include domain knowledge to enhance its value. In addition, we intend to explore the criminal links of border

crossing vehicles and their relationships with gangs and narcotics rings in local law enforcement. This will allow us to map the border crossing patterns of specific gangs and improve the targeting of vehicles.

ACKNOWLEDGMENT

We thank our BorderSafe project partners: Tucson Police Department, Pima County Sheriff's Department, Tucson Customs and Border Protection, ARJIS (Automated Regional Justice Information Systems), San Diego Super Computer Center (SDSC), Department of Homeland Security, and Corporation for National Research Initiatives (CNRI).

REFERENCES

- [1] "National Strategy for Homeland Security," Office of Homeland Security, 2002.
- [2] B. Marshall, S. Kaza, J. Xu, H. Atabakhsh, T. Petersen, C. Violette, and H. Chen, "Cross-Jurisdictional Criminal Activity Networks to Support Border and Transportation Security," in Proc. 7th International IEEE Conference on Intelligent Transportation Systems, Washington D.C., 2004.
- [3] I.-M. A. Chen and D. Rotem, "Integrating Information from Multiple Independently Developed Data Sources," in Proc. 7th International Conference on Information and Knowledge Management, Bethesda, Maryland, 1998.
- [4] R. M. Fano, *Transmission of Information*. Cambridge, MA: MIT Press, 1961.
- [5] K. W. Church and P. Hanks, "Word Association Norms, Mutual Information, and Lexicography," *Computational Linguistics*, vol. 16, pp. 22-29, 1990.
- [6] D. Hindle, "Noun Classification from Predicate-Argument Structures," in Proc. 28th conference on Association for Computational Linguistics, 1990.
- [7] T. Ong and H. Chen, "Updateable PAT-Tree Approach to Chinese Key Phrase Extraction Using Mutual Information: A Linguistic Foundation for Knowledge Management," in Proc. Second Asian Digital Library Conference, Taipei, Taiwan, 1999.
- [8] T. Tao, C. X. Zhai, X. Lu, and H. Fang, "A study of statistical methods for function prediction of protein motifs," *Applied Bioinformatics*, vol. 3, pp. 115-124, 2004.
- [9] D. Weisser and J. Klein-Seetharaman, "Identification of Fundamental Building Blocks in Protein Sequences Using Statistical Association Measures," in Proc. ACM Symposium on Applied Computing, Nicosia, Cyprus, 2004.