

On the Analysis of Illicit Supply Networks using Variable State Resolution-Markov Chains

Jorge Ángel González Ordiano¹, Lisa Finn², Anthony Winterlich², Gary Moloney², and Steven Simske¹

¹ Colorado State University, Fort Collins, CO, USA
{jorge.gonzalez_ordiano,steve.simske}@colostate.edu

² Micro Focus International, Galway, Ireland
{finn,winterlich,gary.moloney}@microfocus.com

Abstract. The trade in illicit items, such as counterfeits, not only leads to the loss of large sums of private and public revenue, but also poses a danger to individuals, undermines governments, and—in the most extreme cases—finances criminal organizations. It is estimated that in 2013 trade in illicit items accounted for 2.5% of the global commerce. To combat illicit trade, it is necessary to understand its illicit supply networks. Therefore, we present in this article an approach that is able to find an optimal description of an illicit supply network using a series of Variable State Resolution-Markov Chains. The new method is applied to a real-world dataset stemming from the Global Product Authentication Service of Micro Focus International. The results show how an illicit supply network might be analyzed with the help of this method.

Keywords: Data Mining · Markov Chain · Illicit Trade

1 Introduction

Illicit trade is defined as the trade in illegal goods and services that have a negative impact on our economies, societies, and environments [12]. Two of the most prevalent forms of illicit trade are counterfeiting and piracy, whose negative effects have been studied by both the OECD and the ICC. The former estimates that in 2013 counterfeiting and piracy accounted for 2.5% of all world imports [10], while the latter assesses that by 2022 counterfeiting and piracy will drain 4.2 trillion dollars from the world economy and put 5.4 million jobs at risk.¹ The consequences of illicit trade go beyond the loss of public and private revenue. Counterfeit medicines, for instance, have caused a large number of malaria and tuberculosis related deaths [6], while counterfeit cigarettes, cd's, etc. have been linked to terrorist organizations [1]. These examples show the danger that illicit trade poses to our communities. Therefore, finding ways to combat this type of trade is of paramount importance.

A possibility for battling illicit trade is through the disruption of its illicit supply networks (ISNs). Different methods on how to achieve this disruption

¹ iccwbo.org/global-issues-trends/bascap-counterfeiting-piracy/, Accessed:07-17-2019

are found in literature. Many articles deal with technologies for distinguishing between licit and illicit goods, such as the works of Dégardin et al. [3], Simske et al. [15], and Meruga et al. [9]. More closely related to the present article are those in which the ISNs are investigated directly. Some examples of this type of articles are shown by Giommoni et al., [5], Magliocca et al. [7], and Triepels et al. [16]. In the first, network analysis of the heroin trafficking networks in Europe is conducted. In the second, a simulation of the response of drug traffickers to interdiction is presented. In the third, international shipping records are used to create Bayesian networks able to detect smuggling and miscoding.

The goal of this article is to identify the locations in which illicit activity is more prevalent. To achieve this goal, we make use of Markov Chains, as they are a type of model that is useful at determining the amount of time that a system (i.e. a supply network) spends on a given state (i.e. a location) [13]. The first step for creating a Markov Chain is to define what the states, or nodes, of the model will be. These states can be defined at different resolution levels, as shown for instance in [8]. In this article, the states represent possible geographic locations within a supply network; which in turn can be defined in terms of countries, regions, continents, etc. Unfortunately, the state description (i.e. the Markov Chain design) that is best at modeling a given system is not immediately clear. To address this issue, we present in this work a new method that optimizes—in terms of a user-defined cost function—the design of a Markov Chain. Notice that the models created via this new method are referred to as Variable State Resolution-Markov Chains (VSR-MCs) to denote the fact that their states are a combination of the various possible descriptions. Furthermore, a real world dataset containing spatio-temporal information of serial code authentications is used to show how the new approach can be used to analyze ISNs. This dataset stems from the Global Product Authentication Service of Micro Focus International. The VSR-MC obtained with this data is then used to compare a licit supply network to its illicit counterpart. The results of this comparison offer insight on the locations in which illicit activity is more prevalent.

The remainder of this article is organized as follows: Section 2 offers preliminary information on Markov Chains. Section 3 shows the new method. Section 4 describes this article’s experiment. Section 5 shows and discusses the obtained results and Section 6 contains the conclusion and outlook of this work.

2 Preliminaries

A Markov Chain can be defined as a discrete time random process $\{X_n : n \in \mathbb{N}_0\}$ whose random variables² only take values within a given state space, i.e. $x_n \in S$ [2]. In this section, a state space—consisting of $K \in \mathbb{N}_{>1}$ different states—is defined as $S = \{s_k : k \in [1, K]\}$. In general, a Markov Chain can be viewed as a Markov Process with discrete time and state space.

² Note that the common notation for random variables is used herein, i.e. random variables are written in uppercase and their realizations in lowercase.

The most important property of a Markov Chain is its lack of “memory”, i.e. the probability of an outcome at time $n + 1$ depends only on what happens at time n [11]. This is better described by the following equation:

$$P(X_{n+1}|X_n = x_n, \dots, X_1 = x_1) = P(X_{n+1}|X_n = x_n) . \quad (1)$$

A Markov Chain is further characterized by its transition probability matrix $\mathbf{P}_n$; a matrix defined as:

$$\mathbf{P}_n = \begin{bmatrix} p_{11} & \cdots & p_{1K} \\ \vdots & \ddots & \vdots \\ p_{K1} & \cdots & p_{KK} \end{bmatrix}_n , \quad (2)$$

with the entries $p_{ij,n}$ representing the probability of transitioning from state s_i to state s_j at time n , i.e. $P(X_{n+1} = s_j|X_n = s_i)$. If the transition probabilities are independent of n (i.e. $\mathbf{P}_n = \mathbf{P}$), the Markov chain is called time homogeneous [11].

Additionally, the probabilities of X_0 being equal to each one of the states can be written in vector form as follows:

$$\boldsymbol{\pi}_0 = [P(X_0 = s_1), \dots, P(X_0 = s_K)]^T = [\pi_{01}, \dots, \pi_{0K}]^T , \quad (3)$$

where $\boldsymbol{\pi}_0$ is the start probability vector and π_{0k} is the probability of X_0 being equal to s_k .

Based on Equations (1), (2), and (3), the probability of a sequence of events in a time homogeneous Markov Chain can be calculated as a multiplication of a start probability and the corresponding p_{ij} values [14]. For instance, the probability of the sequence $\{X_0 = s_1, X_1 = s_3, X_2 = s_2\}$ is given as:

$$P(X_0 = s_1, X_1 = s_3, X_2 = s_2) = \pi_{01} \cdot p_{13} \cdot p_{32} . \quad (4)$$

Interested readers are referred to [2] and [11] for more information on Markov Chains.

3 Variable State Resolution-Markov Chain

The method presented herein offers a novel alternative on how to optimize the design of a Variable State Resolution-Markov Chain (VSR-MC). The main difference between a traditional Markov Chain and a VSR-MC is the way in which the state space is defined. This difference stems from the fact that a state can be defined at different resolution scales, which are referred in this article as scales of connectivity. For example, a geographic location within a supply network can be described at a country or at a continent scale. Based on this idea, we define the state space of a VSR-MC as:

$$\begin{aligned} S &= \{\Phi_G(s_k) : k \in [1, K]\} \\ &= \{s_{G,k'} : k' \in [1, K_G]\} , \text{ with} \\ G &= \{g_{lr} : l \in [1, L], r \in [1, R_l]\} , \end{aligned} \quad (5)$$

where s_k represents the states, G is a set containing the groups (i.e. g_{lr}) in which the states can be clustered, $L \in \mathbb{N}_{>0}$ is the number of scales of connectivity, and $R_l \in \mathbb{N}_{>1}$ is the number of groups within the l^{th} scale. Furthermore, $\Phi_G(s_k)$ is a function that defines $K_G \in \mathbb{N}_{>1}$ new states, which are referred to as $s_{G,k'}$. In other words, $\Phi_G(s_k)$ is defined as follows:

$$\Phi_G(s_k) = \begin{cases} s_k & , \text{ if } s_k \notin G \\ g_{lr} : s_k \in g_{lr} & , \text{ else } . \end{cases} \quad (6)$$

When defining G , it is important to consider that each s_k can only be **contained in either one or none of the groups within the set**. For the sake of illustration, Fig. 1 shows an example of possible high resolution states and their corresponding groups at different scales of connectivity.

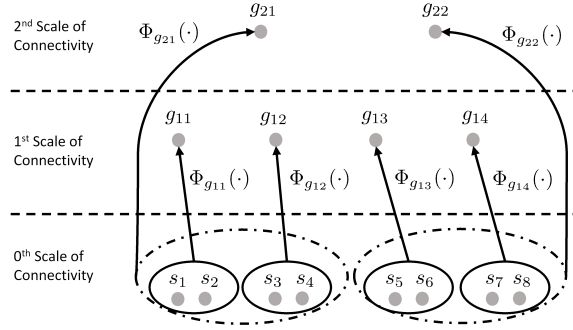


Fig. 1. Example of various states and their corresponding groups

Based on all previous aspects, it is clear that we can create different Markov Chains based on the combination of different states and groups. For instance, consider a case in which four states (i.e. s_1 , s_2 , s_3 , and s_4) can be aggregated in two groups (i.e. $g_{11} = \{s_1, s_2\}$ and $g_{12} = \{s_3, s_4\}$). As shown in Table 1, the possible combinations result in four VSR-MCs with different scales of connectivity.

In general, the number of combinations (i.e. group sets G) that can be obtained with L scales of connectivity is given by the next equation:

$$N_c = 1 + \sum_{l=1}^L (l+1)^{R_l} - l^{R_l} , \quad (7)$$

where N_c is the number of all possible combinations and R_l is again the number of groups within each scale.

After defining the group sets G , the probabilities of the group set-dependent transition matrices ($\mathbf{P}_G$) and start probability vectors ($\boldsymbol{\pi}_{G,0}$) are calculated.

Table 1. Possible state spaces of the Variable State Resolution-Markov Chains with four states s_1, s_2, s_3 , and s_4 , one scale of connectivity, and two groups $g_{11} = \{s_1, s_2\}$ and $g_{12} = \{s_3, s_4\}$. As given by Eq. (7), the number of possible group sets equals four.

G	$\{\}$	$\{g_{11}\}$	$\{g_{12}\}$	$\{g_{11}, g_{12}\}$
S	$\{s_1, s_2, s_3, s_4\}$	$\{g_{11}, s_3, s_4\}$	$\{s_1, s_2, g_{12}\}$	$\{g_{11}, g_{12}\}$

These probabilities are obtained using a dataset containing N sequences of events of the system we want to model. In this article the sequences are described as:

$$\mathbf{x}_m = \{x_{mn} : n \in [0, N_m]\}, \quad (8)$$

in which $\mathbf{x}_m$ is the m^{th} sequence within the dataset, $N_m \in \mathbb{N}_{>0}$ is a value that defines the sequence length, and x_{mn} is one of the realizations forming the sequence.

As mentioned at the beginning, the main goal is to find the scale of connectivity that will optimize the Markov chain architecture. In other words, we are interested in finding the VSR-MC that minimizes a problem-specific cost function $c(\cdot)$. This optimization problem can be described in general as:

$$G_{\text{opt}} = \underset{G}{\operatorname{argmin}} c(G, S, \mathbf{P}_G, \boldsymbol{\pi}_{G,0}, \dots), \quad (9)$$

where G_{opt} represents the optimal group set.

4 Experimental Study

4.1 Data

The dataset used comes from the Global Product Authentication Service (GPAS) of Micro Focus International. GPAS protects products in the marketplace by embedding a URL and unique serial number into a QR code placed on each product. The consumer is encouraged to scan the QR code which can authenticate their purchase in real-time. This dataset contains therefore spatio-temporal information of licit and illicit activity. To be more specific, it contains the authentication results (i.e. “True” or “False”) of 1,725,075 unique serial codes.³ In addition to the authentication, the dataset contains the geographic position (i.e. latitude and longitude) and the time at which each serial code was authenticated. Since many codes have been authenticated several times at different times and places, we assume that a reconstruction of the supply network is possible.

In the present article, we are interested in analyzing licit and illicit serial codes that are authenticated a similar number of times at different geographic locations. Henceforth, the data is preprocessed as follows. First, all entries with

³ The serial codes correspond to five different products. In this article, however, they are not separated by their product type, but are rather investigated as a single group.

missing geographic information, as well as all serial codes that do not change their position are removed from the dataset (i.e. 1,659,726). Afterwards, codes whose authentication result is sometimes “True” and sometimes “False” are also eliminated (i.e. 5,453). Note that the serial codes that have been removed are still of interest, as they can be used in the future for other type of analysis. For instance, serial codes that do not change position could be used to identify hot spots of serial code harvesting, while serial codes that change their authentication can be use to analyze locations in which the original licit codes might have been copied. As mentioned earlier, the serial codes we are considering here are the ones authenticated at different locations. We do this because we are interested in discovering the network architecture, and by inference the distribution channels, of the illicit actors. Finally, serial codes authenticated first and last at the exact same position, as well as those authenticated in more geographic positions than 99% of all serial codes are deleted (i.e. 3,897). The goals of this final step are the removal of serial codes that are suspect of being demos and the elimination of copied serial codes authenticated a huge number of times (i.e. with a clearly different behavior than licit serial codes).

The resulting dataset contains 55,999 unique serial codes, of which 31,989 are authenticated as “True”, while 24,010 are authenticated as “False”.

4.2 Description

The goal of this experiment is to find a VSR-MC able to accurately describe a licit and an illicit supply network. To do so, we create a series of Markov Chains with computed probabilities of state-state transitions for both licit and illicit serial codes. Then, we select the one that is best at classifying illicit activity as the one with the optimal scale of connectivity. To solve this classification problem and to obtain representative results, we create three different training/test set pairs, by randomly selecting—three separate times—50% of the unique serial codes as training set and the rest as test set.

We begin the experiment by defining three different ways in which the location of a serial code can be described, i.e. country, region, or continent. These descriptions are the scales of connectivity of the VSR-MC we are looking to create. Using the given geographic positions, we can easily determine the countries and continents where the serial codes were authenticated. The regions, in contrast, are calculated using a clustering algorithm, i.e. the affinity propagation algorithm [4]. This algorithm clusters the countries of a specific continent based on a similarity measure. The similarity measure we use here is the geographic proximity between the countries’ centroids. For the sake of illustration, Table 2 shows the three scales of connectivity used in this article.

As Eq. (7) shows, the regional and continental descriptions can be used to create a staggering number of possible combinations; whose individual testing would be computationally infeasible. For this reason, the next paragraphs describe a two step alternative to deal with this issue.

Step 1: We begin by finding an optimal VSR-MC using only the regions and continents. In other words, we define the regions as the Markov states s_k

Table 2. Scales of connectivity

Continents	Regions	Countries
Africa	Africa 1	Djibouti, Egypt, Ethiopia, Kenya, Sudan
	Africa 2	Angola, Botswana, Mozambique, Namibia, South Africa, Zambia, Zimbabwe
	Africa 3	Algeria, Libya, Morocco, Tunisia
	Africa 4	Burundi, Congo - Kinshasa, Malawi, Rwanda, South Sudan, Tanzania, Uganda
	Africa 5	Benin, Burkina Faso, Côte d'Ivoire, Ghana, Mali, Niger, Nigeria, Togo
	Africa 6	Cape Verde, Gambia, Guinea, Guinea-Bissau, Liberia, Mauritania, Senegal, Sierra Leone
	Africa 7	Cameroon, Congo - Brazzaville, Gabon
	Africa 8	Comoros, Madagascar, Mauritius, Réunion, Seychelles
Asia	Asia 1	Bangladesh, Bhutan, India, Maldives, Myanmar (Burma), Nepal, Sri Lanka
	Asia 2	Japan, South Korea
	Asia 3	Bahrain, Iran, Kuwait, Oman, Qatar, Saudi Arabia, United Arab Emirates, Yemen
	Asia 4	Cambodia, Hong Kong SAR China, Laos, Macau SAR China, Taiwan, Thailand, Vietnam
	Asia 5	Brunei, Indonesia, Malaysia, Philippines, Singapore
	Asia 6	Armenia, Azerbaijan, Georgia, Iraq, Israel, Jordan, Lebanon, Palestinian Territories, Syria, Turkey
	Asia 7	China, Mongolia, Russia
	Asia 8	Afghanistan, Kazakhstan, Kyrgyzstan, Pakistan, Tajikistan, Turkmenistan, Uzbekistan
Europe	Europe 1	Albania, Bulgaria, Cyprus, Greece, Macedonia, Malta, Montenegro, Serbia
	Europe 2	Iceland
	Europe 3	Moldova, Romania, Ukraine
	Europe 4	France, Portugal, Spain
	Europe 5	Belgium, Denmark, Germany, Ireland, Luxembourg, Netherlands, United Kingdom
	Europe 6	Åland Islands, Finland, Norway, Sweden
	Europe 7	Austria, Bosnia & Herzegovina, Croatia, Czechia, Hungary, Italy, Monaco, Slovakia, Slovenia, Switzerland
	Europe 8	Belarus, Estonia, Latvia, Lithuania, Poland
North America	North America 1	United States
	North America 2	Canada
	North America 3	Mexico
	North America 4	Bahamas, Curaçao, Dominican Republic, Haiti, Jamaica
	North America 5	Barbados, British Virgin Islands, Guadeloupe, Martinique, Puerto Rico, Sint Maarten, Trinidad & Tobago
	North America 6	Costa Rica, El Salvador, Guatemala, Honduras, Nicaragua, Panama
Oceania	Oceania 1	Australia, Fiji, New Zealand, Samoa
	Oceania 2	Northern Mariana Islands, Palau, Papua New Guinea
South America	South America 1	Bolivia, Brazil, Paraguay
	South America 2	Colombia, Ecuador, Peru
	South America 3	Argentina, Chile, Uruguay
	South America 4	French Guiana, Guyana, Venezuela

and the continents as the groups at the first—and only—scale of connectivity, i.e. $g_{1r} \in \{\text{Africa, Asia, North America, South America, Oceania}\}$. Using these groups and Eq. (5), we define a total of 64 different state spaces S .

Afterwards, we calculate for each state space and each available training set the probabilities that are necessary to construct a Markov Chain. To be more specific, for each state space two start probability vectors and two transition probability matrices are calculated, for licit and illicit serial codes, respectively. These probabilities are based on the trajectories that are described herein as a sequence of geographic positions in which a serial code has been authenticated. Based on Eq. (8), the trajectories can be described as:

$$\mathbf{x}_m^\alpha = \{x_{mn}^\alpha : n \in [0, N_m]\} : \alpha = \{\text{licit, illicit}\}, \quad (10)$$

where α indicates if the trajectory corresponds to a licit or an illicit serial code.

Notice that we assume the Markov Chains to be homogeneous. Therefore, all sequences within the training sets that have an $N_m > 1$ are divided in N_m sequences of two realizations each. With these new set of sequences as well as Eq. (5), the start probabilities can be calculated:

$$\pi_{G,0i}^\alpha = \frac{1}{M^\alpha} \sum_{m=1}^{M^\alpha} I(x_{m0}^\alpha = s_{G,i}) : \alpha = \{\text{licit, illicit}\}, \quad i \in [1, K_G], \quad (11)$$

where $\pi_{G,0i}^\alpha$ is the probability of a sequence starting at state $s_{G,i}$, M^α represents the number of available sequences (licit or illicit), and $I(\cdot)$ is a function that equals one if its condition is fulfilled and equals zero otherwise. Thereafter, the elements of the transition matrices can be obtained using Bayes's rule:

$$p_{G,ij}^\alpha = \frac{1}{M^\alpha \pi_{G,0i}^\alpha} \sum_{m=1}^{M^\alpha} I(x_{m0}^\alpha = s_{G,i} \cap x_{m1}^\alpha = s_{G,j}) \quad (12)$$

$$: \alpha = \{\text{licit}, \text{illicit}\}, i, j \in [1, K_G],$$

with $p_{G,ij}^\alpha$ being the probability of transitioning from state $s_{G,i}$ to $s_{G,j}$.

So, using the previous values we can determine—in every state space—the probability of a serial code sequence if we assume it to be licit or illicit, i.e. $P_G(\mathbf{x}_m|\alpha) : \alpha = \{\text{licit}, \text{illicit}\}$. Thereafter, we can use the resulting probabilities to classify a serial code as illicit if $P_G(\mathbf{x}_m|\text{illicit}) \geq P_G(\mathbf{x}_m|\text{licit})$.

The method described previously is used to classify all codes within the test sets. Afterwards, the classification results are evaluated using the weighted F-Score, i.e.:

$$F_{\beta,G} = (1 + \beta^2) \frac{Q_{p,G} Q_{r,G}}{\beta^2 Q_{p,G} + Q_{r,G}}, \quad (13)$$

where $Q_{p,G}$ represents the precision, $Q_{r,G}$ is the recall, and β is a parameter that defines which of the former values is weighted more strongly. The value of β is set equal to 0.5 to give precision two times more importance than recall. This is done, since we are more interested in correctly identifying illicit serial codes (precision) than we are in flagging every possible one (recall).

After finishing the evaluation on each test set, the mean value and variance of the weighted F-Score are calculated, i.e. $\bar{F}_{\beta,G}$ and $\sigma_{\beta,G}^2$, respectively. With these values, the optimization problem described in Eq. (9) can be redefined as:

$$G_{\text{opt}} = \underset{G}{\operatorname{argmin}} \gamma (1 - \bar{F}_{\beta,G}) + (1 - \gamma) \sigma_{\beta,G}^2 : \gamma \in [0, 1]. \quad (14)$$

Notice that in this article the parameter γ is set equal to 0.5 to give both terms of the cost function an equal weight and to make the cost function less sensitive to noise. Solving Eq. (14) results in an optimal VSR-MC whose state space S (defined by G_{opt}) might be a combination of regions and continents.

Step 2: If the number of regions within the state space S is greater than zero, we can conduct an additional experiment to test if a country level description of the regions improves our modeling of the supply network. Note that our experiment uses a forward selection to reduce the number of combinations that need to be tested. We first create new state spaces by individually separating the regions within S into their corresponding countries. For instance, if S contains 6 regions we obtain 6 new state spaces. Afterwards, we test if some of these new state spaces result in a Markov Chain with a cost (cf. Eq (14)) that is lower than the one currently consider optimal. If so, we define the VSR-MC with the lowest cost as the new optimal one and its state space as the new optimal state space S . Afterwards, we repeat the previous steps again until none of the new

Markov Chains result in a better cost or until there are no more regions within the state space. The result of this process is a VSR-MC with states that could stem from all of our available scales of connectivity (i.e. countries, regions, and continents). For the sake of simplicity, we will refer to the group set that maps the individual countries to the state space of this new optimal VSR-MC also as G_{opt} . It is worth noting, that since we are not testing all possible state spaces, the solution of our method may not be the global optimum. Nevertheless, we still consider our approach of dividing one region at a time to be acceptable. There are two main reasons for this: (i) we are able to improve the overall cost function testing only a small subset of all combinations; and (ii) we are able to increase the resolution of our network description, something that may improve our understanding of how the network operates.

After finding the best VSR-MC, we use all available data to recalculate the probabilities of the transition matrices to analyze the differences between the licit and the illicit supply networks with more detail. The analysis consists in calculating the limiting distributions of the licit and illicit transition matrices. These describe the probabilities of authenticating the serial codes at the different locations if we observe our system (i.e. the supply network) over a long period of time. In other words, these values can be interpreted as estimates of the amount of time that licit or illicit serial codes will spend on the different locations. Therefore, a comparison of licit and illicit limiting distributions will allow us to estimate the locations where we expect illicit serial codes to spend more time. The comparison is based on a relative difference that is defined in this article as:

$$\Delta\pi'_{s_{G_{\text{opt}},i}} = \frac{\pi_{s_{G_{\text{opt}},i}}^{\text{licit}} - \pi_{s_{G_{\text{opt}},i}}^{\text{illicit}}}{\pi_{s_{G_{\text{opt}},i}}^{\text{licit}}}, \quad (15)$$

where $\pi_{s_{G_{\text{opt}},i}}^{\text{licit}}$ is the licit limiting distribution value of $s_{G_{\text{opt}},i}$, $\pi_{s_{G_{\text{opt}},i}}^{\text{illicit}}$ represents the illicit limiting distribution value of $s_{G_{\text{opt}},i}$, and $\Delta\pi'_{s_{G_{\text{opt}},i}}$ is the relative difference for state $s_{G_{\text{opt}},i}$.

After estimating the relative differences, we can test the difference between our approach and a simple descriptive analysis. This test consists in comparing the $\Delta\pi'_{s_{G_{\text{opt}},i}}$ values to benchmark relative differences (BRDs) calculated using descriptive statistics. To be more specific, the BRDs are also obtained with Eq. (15), but instead of using the limiting distribution values, we use the actual percentage of true and false authentications on the given states.

5 Results and Discussion

The results obtained on the three separate test sets by the VSR-MCs with only regions and continents as scales of connectivity (cf. Sec. 4.2; Step 1) are depicted in Fig. 2.

The first thing we notice when looking at Fig. 2 is that the standard deviations are relatively small. This not only means that the results on all test sets are similar, but also that our mean estimates are quite accurate, as the standard

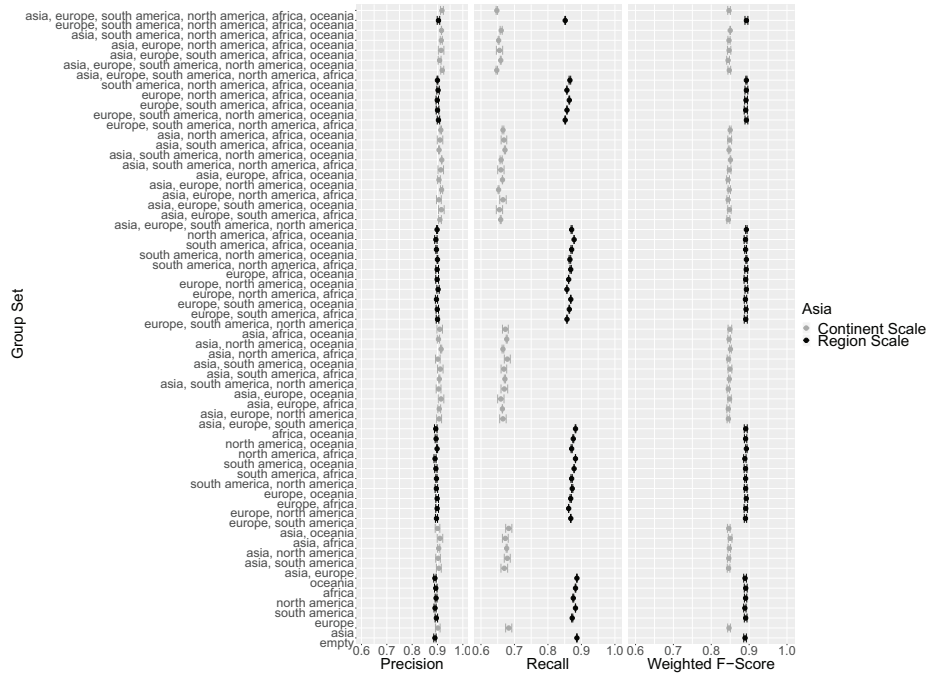


Fig. 2. Group set-dependent mean values and standard deviations of the precision, recall, and weighted F-score (cf. Eq. (13)) obtained on the three separate test sets

error of the mean is directly proportional to the standard deviation. In addition, Fig. 2 also shows that the precision does not appear to change when modifying the state space; as it is consistently around 90%. In contrast, the use of different group sets divides the recall in two distinct groups with different recall values; the first between 60 and 70% and the second between 80 and 90%. The decrease in recall is caused by considering Asia as a continent instead of looking at its individual regions. In other words, individual networks between Asian regions seem to play an important role in the accurate modeling of licit and illicit supply networks. Due to the recall, the weighted F-Score is also dependent on Asia being modeled as a single state or as individual regions. This result, i.e. that the scale of connectivity affects the quality of the supply network models, supports the use of this article’s method (cf. Sec. 3). Therefore, we use Eq. (14) and the obtained weighted F-Scores to determine the scale of connectivity that will best describe the licit and illicit supply networks.

According to Eq. (14), the optimal VSR-MC is the one with states representing the continents of Africa, Europe, North America, South America, and Oceania, as well as the individual regions of Asia, i.e. $G_{\text{opt}} = \{\text{Africa, Europe, North America, South America, Oceania}\}$. This result shows again the importance that Asia appears to play in the accurate modeling of the supply networks.

After finding the best VSR-MC, we can identify the regions that, when divided, improve our model (cf. Sec. 4.2; Step 2). Our method concludes that our description of the licit and illicit supply networks improve if we consider five of the eight Asian regions (i.e. Asia 1, Asia 5, Asia 6, Asia 7, and Asia 8) as individual countries. Therefore, we redefine the optimal group set as $G_{\text{opt}} = \{\text{Asia 2, Asia 3, Asia 4, Africa, Europe, North America, South America, Oceania}\}$. This group set defines a new VSR-MC with a state space that combines the three scales of connectivity we considered in this article. Lastly, it is important to mention, that the cost of this new optimal VSR-MC (i.e. 0.047) is not only lower than the one obtained when Asia is divided purely into regions (i.e. 0.054), but also than the one obtained when Asia is divided purely into countries (i.e. 0.048).

Once the optimal scale of connectivity, given any limitations of our process, has been found, we recalculate the Markov Chain probabilities with all available data and use Eq. (15) to identify the states in which illicit serial codes are more prevalent. It is important to mention that having some of the countries as states results in the transition matrices having absorbing states; a type of state that complicates the calculation of the limiting distributions. Therefore to calculate the limiting distributions, we first group those countries with the “less absorbing” countries within their region. In this context, “less absorbing” refers to countries whose rows in their transition matrices have the least number of zeros when compared to all other countries within their region.

The relative differences $\Delta\pi'_{s_{G_{\text{opt}},i}}$ (cf. Eq. (15)), the number of times a serial code is authenticated, and the benchmark relative differences (BRDs) are all contained in Table 3.

When looking at Table 3, we notice a state with a relative difference of minus infinity (i.e. Mongolia) and another with a relative difference of one (i.e. Tajikistan). This means that in those locations only illicit or only licit serial codes were authenticated. Even though these types of results might be interesting, they will not be investigated further, as the number of authentications in those locations is extremely low.

Table 3 also shows that the countries that would have formed absorbing states are locations in which serial codes are authenticated a small number of times, specially compared to the number of authentications within the “less absorbing” countries they are grouped with (i.e. India, Malaysia, Pakistan, and Turkey). Henceforth, we can safely assume that the “less absorbing” countries are the ones responsible for the relative differences obtained. Furthermore, the results in Table 3 show that there are several states in which illicit serial codes appear to spend more time than licit ones. These states are the ones with a negative relative difference (cf. Eq. (15)) and are further referred to as “critical” states. The fact that most of these critical states are countries within the regions selected by our forward selection algorithm (cf. Sec. 4.2; Step 2), speaks in favor of our approach.

As Table 3 shows, Turkey is the most “critical” state, as its relative difference estimates that illicit serial codes will spend close to 1200% more time there than their licit counterparts. This is an extreme result that needs to be

Table 3. Number of authentications, relative limiting distribution difference $\Delta\pi'_{s_{G_{opt},i}}$, and benchmark relative difference (BRDs) of the states forming the optimal Variable State Resolution-Markov Chain; the absorbing states and their realization are shown in parentheses next to the the name and realizations of their corresponding “less absorbing” countries

$s_{G_{opt},i}$	# of Authentications	$\Delta\pi'_{s_{G_{opt},i}}$	BRD
Mongolia	1	-Inf	-Inf
Turkey (Jordan, Armenia)	42064 (53, 5)	-11.69	-12.07
Georgia	16	-2.91	-1.72
Singapore	189	-2.51	-0.69
Syria	22	-2.41	-0.59
Europe	20621	-0.49	-1.87
Kazakhstan	114	-0.32	0.47
Pakistan (Kyrgyzstan, Turkmenistan, Uzbekistan)	1944 (1, 1, 5)	-0.32	-0.02
China	19923	-0.20	0.80
Sri Lanka	298	-0.07	0.42
North America	10991	-0.01	0.48
South America	24590	0.00	0.50
Asia 3	3606	0.25	0.51
India (Nepal, Maldives)	13669 (38, 29)	0.37	0.64
Russia	4475	0.42	0.59
Philippines	283	0.52	0.78
Bangladesh	282	0.53	0.51
Bhutan	279	0.53	-0.53
Africa	11756	0.56	0.58
Afghanistan	445	0.58	0.50
Asia 4	1815	0.65	0.73
Malaysia (Brunei)	1898 (16)	0.69	0.75
Indonesia	318	0.70	0.77
Palestinian Territories	113	0.71	0.28
Oceania	644	0.76	0.76
Israel	363	0.87	0.23
Asia 2	654	0.89	0.49
Lebanon	56	0.92	0.80
Azerbaijan	457	0.92	0.95
Iraq	32	0.98	0.91
Myanmar	92	0.99	0.98
Tajikistan	3	1.00	1.00

investigated further, for instance by identifying the reasons behind this outcome and/or by finding out if Turkey is again a critical state when looking at illicit activities, such as serial code harvesting. The critical states with the next three lowest relative differences are Georgia, Singapore, and Syria. There the limiting distribution values of an illicit serial code are between 200 and 300% higher than those of a licit one. However, we can also observe that the number of authentications occurring on those locations is quite low in comparison to other places. Henceforth, a further investigation of those locations may not be of extreme importance. In addition to the results mentioned above, there are several “critical” states with relative difference that can still be considered high, i.e. between 20 and 50%. Within these states, Europe and China are the ones with a considerably larger amount of authentications. Therefore, a more in depth study of these two locations could be interesting for future related works.

It is also important to mention that a state having a relative difference close to zero does not mean that it is free of illicit activity. For instance, Sri Lanka, North America, and South America have relative differences of just -0.07, -0.01, and 0, respectively, meaning that their limiting distribution values for licit and illicit serial codes are almost the same. In other words, states whose relative differences are close to zero may have a similar rate of licit and illicit activity and thus should be investigated further.

Lastly, we observe that some relative difference values vary significantly or do not agree to those calculated using simple descriptive statistics (i.e. the BRDs). For example, though China has the relative difference of a so-called “critical” state, its BRD is clearly above zero. Similarly, North and South America have BRDs that indicate more licit than illicit authentications, while their relative differences estimate instead similar rates. Note that though we are aware about the interesting results obtained for Bhutan (its BRD and its relative difference are complete opposites), no further investigation and analysis were conducted given its relatively small number of authentications. In general, the results obtained in this work demonstrate that modeling the spatio-temporal information of a supply network (as we do with our approach) leads to conclusions that are different from those obtained through a simple descriptive analysis. Furthermore, since our method models the behavior of the illicit supply network as a whole, we can argue that it is better suited at combating illicit trade than a descriptive analysis.

6 Conclusion and Outlook

This article presents a new approach for describing illicit supply networks based on Variable State Resolution-Markov Chain (VSR-MC) models. These type of models stem from the idea that a location within a supply network can be described at different scales of connectivity (e.g., countries, regions, continents).

The new method described herein is divided in two main steps. The first step creates a series of VSR-MCs that describe the same network using different state spaces, while the second uses a user-defined cost function to select the VSR-MC that best describes the network. The new method is applied to a dataset containing spatio-temporal information of licit and illicit activity. This dataset comes from the Global Product Authentication Service of Micro Focus International and contains information of the time and place in which licit and illicit serial codes have been authenticated. Applying our new method to this dataset results in Markov Chain models of the licit and illicit supply networks. The comparison of both networks enables us to ascertain the geographic locations in which illicit serial codes are expected to spend more time than their licit counterparts.

Even though this article shows a promising approach for analyzing illicit supply networks, there are still a number of aspects that have to be studied in future related works. For instance, in this article all scales of connectivity stem from grouping the countries based on their geographic proximity; therefore, future works should investigate if better descriptions of the illicit supply networks can be obtained by grouping the countries based on e.g., their number of free trade agreements, their culture, or their language. Such a study will allow us to better identify the aspects that drive illicit supply networks. Additionally, we should also use the method described herein to compare networks stemming from different forms of illicit trade, such as counterfeiting, serial code harvesting, and human trafficking. A comparison like this will enable us to identify both sim-

ilarities and differences between different types of illicit trade. Finally, future works should also investigate the use of n^{th} order and non-homogeneous Markov Chains.

Acknowledgments: Jorge Ángel González Ordiano acknowledges the support given by the NSF EAGER grant with the abstract number 1842577, “Advanced Analytics, Intelligence and Processes for Disrupting Operations of Illicit Supply Networks”.

References

1. Bindner, L.: Illicit trade and terrorism financing. Centre d’Analyse du Terrorisme (CAT) (2016)
2. Collet, J.F.: Discrete Stochastic Processes and Applications. Universitext, Springer, Cham, Switzerland (2018)
3. Dégardin, K., Guillemain, A., Klespe, P., Hindelang, F., Zurbach, R., Roggo, Y.: Packaging analysis of counterfeit medicines. *Forensic Science International* **291**, 144 – 157 (2018).
4. Frey, B.J., Dueck, D.: Clustering by passing messages between data points. *Science* **315**(5814), 972–976 (2007)
5. Giommoni, L., Aziani, A., Berlusconi, G.: How do illicit drugs move across countries? A network analysis of the heroin supply to Europe. *Journal of Drug Issues* **47**(2), 217–240 (2017)
6. Mackey, T.K., Liang, B.A.: The global counterfeit drug trade: Patient safety and public health risks. *Journal of pharmaceutical sciences* **100**(11), 4571–4579 (2011)
7. Magliocca, N.R., McSweeney, K., Sesnie, S.E., Tellman, E., Devine, J.A., Nielsen, E.A., Pearson, Z., Wrathall, D.J.: Modeling cocaine traffickers and counterdrug interdiction forces as a complex adaptive system. *Proceedings of the National Academy of Sciences* **116**(16), 7784–7792 (2019)
8. Meidani, H., Ghanem, R.: Multiscale markov models with random transitions for energy demand management. *Energy and Buildings* **61**, 267 – 274 (2013).
9. Meruga, J.M., Cross, W.M., May, P.S., Luu, Q., Crawford, G.A., Kellar, J.J.: Security printing of covert quick response codes using upconverting nanoparticle inks. *Nanotechnology* **23**(39), 395201 (2012)
10. OECD: Governance Frameworks to Counter Illicit Trade (2018).
11. Privault, N.: Understanding Markov Chains: Examples and Applications. Springer Undergraduate Mathematics Series, Springer Singapore, Singapore (2013)
12. Shelley, L.I.: Dark Commerce: How a New Illicit Economy is Threatening our Future. YBP Print DDA, Princeton University Press, Princeton, New Jersey (2018)
13. Simske, S.: Meta-Analytics: Consensus Approaches and System Patterns for Data Analysis. Morgan Kaufmann (2019)
14. Simske, S.J.: Meta-Algorithmics: Patterns for Robust, Low Cost, High Quality Systems. Ebook Central (EBC), IEEE Press, Wiley, Chichester, West Sussex, United Kingdom (2013)
15. Simske, S.J., Sturgill, M., Aronoff, J.S.: Comparison of image-based functional monitoring through resampling and compression. In: Proc., IEEE International Geoscience & Remote Sensing Symposium (2009)
16. Triepels, R., Daniels, H., Feelders, A.: Data-driven fraud detection in international shipping. *Expert Systems with Applications* **99**, 193–202 (2018)