Б Ъ Л Г А Р С К А А К А Д Е М И Я Н А Н А У К И Т Е

ИНСТИТУТ ПО ИНФОРМАЦИОННИ И КОМУНИКАЦИОННИ ТЕХНОЛОГИИ

Иван Иванов Благоев

МЕТОДИ И СРЕДСТВА ЗА АНАЛИЗ НА ДАННИ В ИНФОРМАЦИОННИ СИСТЕМИ С ИЗПОЛЗВАНЕ НА ВРЕМЕВИ РЕДОВЕ

А В Т О Р Е Ф Е Р А Т

НА ДИСЕРТАЦИЯ

за придобиване на образователната и научна степен „доктор“

по докторска програма „Информатика“

професионално направление 4.6. “Информатика и компютърни науки“

Научен ръководител: доц. д-р Татяна Атанасова

София, 2021 г.

Дисертацията е обсъдена и допусната до защита на разширено заседание на секция „Моделиране и оптимизация“ на ИИКТ-БАН, състояло се на .....

Дисертационният труд съдържа 125 страници, 33 фигури, 1 таблица и 122 литературни източника.

Защитата на дисертацията ще се състои на ..... от ..... часа в зала ..... на блок 2 на ИИКТ-БАН на открито заседание на научно жури в състав:

1.
2.
3.
4.
5.

Материалите за защитата са на разположение на интересуващите се в стая ..... на ИИКТ-БАН, ул. „Акад. Г. Бончев“, бл. 2.

Автор: Иван Иванов Благоев

Заглавие: МЕТОДИ И СРЕДСТВА ЗА АНАЛИЗ НА ДАННИ В ИНФОРМАЦИОННИ СИСТЕМИ С ИЗПОЛЗВАНЕ НА ВРЕМЕВИ РЕДОВЕ

Увод

Напредъкът в технологиите е толкова очевиден, че може само да се спомене, без нужда от фактологично описание. В това отношение, значителна разлика от последно време е силно експанзиращата дигитална трансформация. Поради COVID-19 заплахата за човешкото здраве, скоростта на навлизане на технологиите в нашият живот силно се ускори, което води до тотална промяна в множество дейности, а в следващите години ще се забелязва още по-силно, когато човечеството се трансформира и адаптира към този нов начин на живот.

Всичкото споменато до тук, води със себе си и до много напълно нови за науката и неизследвани до сега процеси. Събирането и обработката на времеви поредици и големи данни ще се разшири с проникване и в новите процеси. Нуждата от изследване и нови открития ще е решаваща за развитието на науката и технологиите в следващите години. За това разработката на нови методи и средства за изследванията с времеви редове и обработка на големи данни и е изключително важна и ще бъде основен инструмент за изследванията и развитието на науката и технологиите в бъдеще.

Настоящият дисертационен труд, чрез изследвания с времеви редове допринася за постигане на по-добри резултати при методи за прогнозиране на финансови инструменти, обработката на големи данни и подобряване на криптографията и киберсигурността.

Цел и задачи на дисертацията

Целта на настоящата дисертация е да се разработят нови методи и средства за анализ на данни в информационни системи с използване на времеви редове.

За тази цел се дефинират следните задачи:

1. да се разработи метод за анализ и предсказване на ценови движения във финансовата област с използване на времеви редове;
2. да се предложи алгоритъм за обучение на изкуствени невронни мрежи при прогнозиране на финансови времеви редове;
3. да се предложат решения за повишаване на криптографската защита в информационните системи чрез прилагане на методи за анализ на времеви редове;
4. да се проведат експериментални изследвания за верификация на предложените методи за повишаване на криптографска защита при решаване на задачите за осигуряване на киберсигурността.

5 Да се разработят програмни методи за преодоляване на проблеми при работа с големи обеми от данни във времеви редове.

Структура на дисертацията

Дисертационният труд е структуриран в четири глави.

В **първа** глава е направен преглед на актуалните теми в областта на науката на данните, особено когато тези данни се представят като времеви редове. Мотивирана е необходимостта от разработване на нови методи и средства за анализ на данни в информационни системи с използване на времеви редове.

Във **втора** глава са представени разработените методи за изследване и прогнозиране на финансовите времеви редове с използване на различни математически апарати.

В **трета** глава са описани разработените решения за осигуряване на криптографска защита при предоставяне на информационни услуги чрез изследване на генератори на случайни числа, представляващи поредици от времеви редове. Представено е практическото приложение на предложените подходи за обезпечение на киберсигурността. Показани са реалните резултати от проведените тестове, доказващи успешното решаване на поставените задачи.

В **четвърта** глава преодоляването на проблеми при работа с големи масиви от данни и ограничени компютърни ресурси при изследване на времеви редове е направено с разработените софтуерни подходи и със средствата на език за програмиране R.

В **Заключението** е представено резюме на получените резултати от разработката. Определени са насоки за бъдещи изследвания и развитие. Представен е списък с научни публикации по темата и забелязани цитирания.

Дисертационният труд съдържа 125 страници, 33 фигури, 1 таблица и 122 литературни източника.

Глава 1. Анализ на състоянието на изследванията.

Ако се погледне в околния свят през окото на технологиите, първото което би впечатлило всеки специалист е колко много данни са това. Това е страничен ефект от масовата дигитална трансформация и автоматизацията (Wang, 2020), оставяйки цифрова следа от изпълнението на реалния процес. Тези цифрови следи отразяват случващото се в реалния свят и позволяват задълбочен анализ на основните процеси. Динамичните времеви редове в комуникациите, технологии, бизнеса идват в резултат

на измерване на характеристики от технически, природни, социални, икономически и други системи (Mikalef, 2020), (Ciampi, 2020).

1.1 Времеви редове

Времевите редове представляват редици от данни, събрани на равни или неравни интервали от време. Основна характеристика на времевия ред е, че всяка следваща стойност е в зависимост от предходните стойности. Тази зависимост може да бъде, както много сложна, така и относително проста. Понастоящем много методи за прогнозиране, които действат като ефективни инструменти, са широко приети за оценка и анализ на данни от модели на времеви редове. От тях, най-често използваният модел е интегриран метод на авторегресия със сезонен компонент (SARIMA - Seasonal ARIMA), който по същество принадлежи към линеен модел. Но на практика при решаване на различни задачи в информационни системи най-често срещаното е, че процесът на генериране на данни е силно нелинеен и прогнозите получени с тези модели, не позволяват да се достигне до точните резултати (Martínez-Acosta, 2020).

1.2. Приложение на времеви редове върху финансови инструменти

Пазарните ценови движения се описват чрез времеви редове и са предмет на анализ от финансисти, икономисти и пазарни стратези. Видове финансови анализи, които към настоящия момент се използват за анализ на финансови инструменти:

- Фундаменталният - основава на анализиране на събитията, случващи се по света и касаещи финансовите и стокови пазари (Wafi, 2015);
- Техническият анализ – основава се предимно на статистически методи с изчисления върху времевите редове. Позволява методите за прогнозиране да бъдат описани чрез статистически средства и математически алгоритми (Plummer, 1991), (Scott, 2016).

1.2.1 Невронни мрежи

Подходите за изследване на времеви редове могат да бъдат разделени на две категории: статистически методи и изчислителна интелигентност. Статистическите методи изследват зависимости между изходните и съответните фактори след изучаване на минали данни, докато другата група методи имитира човешкия начин на мислене и логическо заключение, за да придобие знания от миналия опит (като изкуствени невронни мрежи) и да предвиди бъдещи стойности (Atanasova, 2017). Изкуствените невронни мрежи (ANN) се използват в различни научни и ежедневни задачи. Обикновено ANN се представят като претеглен насочен граф и има много различни

конфигурации на тази схема. В най-простия случай това е многослоен персептрон. Времеви редове са атрактивни за изследвания с изкуствени невронни мрежи (Tomov, 2016).

1.3. Приложение на времеви редове при криптографията и кибер сигурността

Изпълнението на изискванията за киберсигурност е предпоставка за безопасността и сигурността на ИТ инфраструктурите, цифровите ресурси и защитата на личните данни. В нейният фундамент е криптографията, която осигурява редица процеси, като идентификация, удостоверяване, кодиране, автентикация, потвърждение за състояние на процеси и данни и др. Основният корен на криптографията са случайните числа, като в най-честия случай за съвременните нужди на криптографията се използват два вида генератори на произволни числа:

- Генератор на случайни числа (RNG);
- Псевдо генератор на случайни числа (PRNG).

Традиционните мерки за RNG са предимно обобщена статистика, отнасяща се до отклонения от математическата случайност. За да се подпомогне проверката на качеството на генератор на случайни числа, може неговият изход да се запише във времеви ред и данните да бъдат подложени на специализирани математически анализи.

1.4 Изводи

В резултат на направените изводи следва да се обобщи, че изследванията на времеви редове в различни области и приложения се нуждаят от разработка на специфични методи и средства за постигане на конкретните цели.

Глава 2. Методи за изследване и прогнозиране на финансовите времеви редове

В тази глава е разгледано изследване върху широко разпространения индикатор Моментум, който принадлежи към групата на осцилаторите. Изчисляването му се базира на математически апарат за обработка на времеви редове. В дисертацията се цели подобряването на неговата ефективност.

2.1.1 Осцилатор Моментум (Momentum Oscillator)

Моментум е основен осцилатор, който показва дали ценовата тенденция се ускорява, забавя или се движи със същата скорост. Той обикновено достига максималната си стойност преди върха на цените и минимума преди дъното на спада.

Функцията на този осцилатор е да отчита ускорението на ценовата тенденция. При изтощаване на текущата тенденция и наличие на вероятност от промяна на същата, Моментум дава сигнал за дивергенция. Това е момент при който цената продължава да се движи в посока на тенденцията, но стойностите на Моментум намаляват при възходяща ценова тенденция или се повишават при низходяща ценова пазарна тенденция. Като потвърждение на сигналите за дивергенция на Моментум в изследването са включени фигурни формации от техническия анализ за комбиниране и потвърждение на текущия ценови обрат. В конкретния пример на фиг.2.2. Моментум и сигналът за дивергенция 1C потвърждава предстоящ ценови пазарен обрат, чрез множествен връх с 1D.

Фиг. 2.2. Дивергенция на Моментум при пазарен тренд EUR/USD на дневна база (исторически данни от Forex пазарът)

2.1.2 Слабости при анализ на пазарния тренд чрез Моментум

В изследването обаче се вижда, че има случаи при които Моментум може да направи изключение и да не отчете дивергенция при завършване на текуща пазарна тенденция, което е изобразено с 1C на фиг.2.3.

Фиг. 2.3. Валутна двойка USD/CAD на дневна база (исторически данни от Forex пазарът)

Според изнесения пример от реалния форекс пазар, стойността на Моментум начертава по-висок връх даже от предходния, но цената след това прави значителна корекция от около 60% без да е на лице дивергенция при осцилатора. Дори напротив, според сигналът, текущата ценова тенденция се потвърждава с ускорението на осцилатора. Въпросът, който вълнува изследването следователно е, дали може да се подобри точността на осцилатора Моментум?

2.1.3 Метод за повишаване точността на Моментум

В този дисертационен труд се предоставя нетрадиционен метод за получаване на сигнал за пазарен обрат, а именно разработеният **MA Volatility Indicator**. Базира се на нетрадиционен начин на използване на индикатор от тип пълзяща средна линия Moving Average (MA). Индикаторът MA се разделя на два под вида:

- Simple moving average (SMA):

  За пресмятане на SMA, се използва времеви ред, при който се сумират данните на последните периоди (*t*), където например *t*=10 за 10 дена, според времевата рамка (може да бъде различна стойност, по избор). След това се дели на броя *t* периодите. Такова пресмятане се прави за всеки един бар за период от графиката. Формулата за SMA е, както следва:

  $$SMA_t = \sum_{n=1}^{t} price_n / t$$

- Exponential moving average (EMA):

С цел да се намали изоставащия ефект на SMA, ползващите технически анализ често предпочитат Exponential Moving Average (EMA). Те намаляват изоставането чрез добавяне на нови стойности върху най-новите цени, зависещи от дължината на МА. Най-кратката EMA ще е с по-голяма стойност, отколкото ще бъде приложена за повечето МА. Формула за пресмятане на EMA:

$$X = K * (C - P) + P,$$

където X – настояща EMA, C – настояща цена, P – EMA от предния период (за пресмятането на първия период се използва стойност от SMA), K – изглаждащ коефициент.

Изглаждащият коефициент прилага подходящ коефициент към по-новите цени, които са свързани с предходните цени на EMA. Формула за изглаждащия коефициент:

$$K = 2 / (1+ N),$$

където N – брой на предходните EMA цени.

Конвенционален подход за търговия чрез МА е на по-висока времева рамка цената да не пресича МА, като при пазарна корекция достигането на МА от пазарната цена се счита за силна подкрепа за текущата тенденция. При пробив на цената на МА се приема, като сигнал за обрат, а при отскачане, като сигнал за потвърждение на текущата пазарна тенденция. Другият метод е анализ с повече от една МА, като всичките МА са с различна скорост. При пробив или отскачане на по-бързата МА към по-бавната МА, се тълкува аналогично за сигнал за потвърждение или обрат в текущата тенденция.

Разработеният в дисертация метод MA Volatility Indicator разчита на определяне на екстремни стойности за отдалечаване на цената от МА, на база на което да се определи сантимента на участниците на пазара към текущия момент. При екстремно високи стойности на отдалечаване на цената от МА по посока на текущата тенденция, трябва да се счита, че това е сигурен сигнал за предстоящ обрат или дълбока корекция на текущата тенденция. Интересното е, че това явление се наблюдава добре в моменти, когато Моментум не дава сигнали за дивергенция и край на ускорението на текущата ценова тенденция.

Фиг.2.8 Комбиниране на Моментум с предлаганото решение (исторически данни от Forex пазарът USD/CAD)

На фигура 2.8 методът MA Volatility Indicator е приложен и комбиниран със стойностите на Моментум, като данните за симулацията са реални исторически от форекс пазара. На фигура 2.8 ясно личи, че Моментум, изобразен със зелена линия, не отчита по-ниски стойности на последния пазарен връх. Това е подчертано с правата червена линия на неговата тенденция. В същото време личи и най-високата на отдалечаване на цената от пурпурната линия на MA със стойност 554 пипса.

В заключение може да се каже, че Моментум е един ефективен осцилатор, който е станал част от множество автоматизирани системи и стратегии за търговия на финансовите пазари. Но разработения в дисертацията методът MA Volatility Indicator, успява да подобри точността на прогнозиране. Следователно, той би могъл да се прилага, както при автоматизирани системи, така и при анализа на пазарните тенденции и от човек.

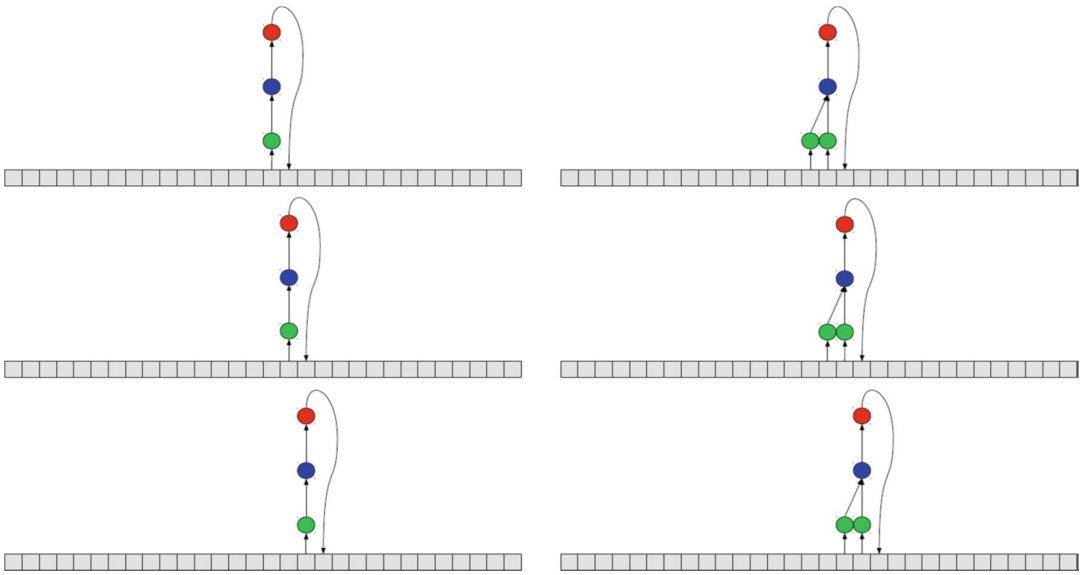
2.2 Прогнозиране на финансови времеви редове чрез невронни мрежи

Многослойният персептрон е най-често използваният вид на изкуствени невронни мрежи, който може да се представи като ориентиран претеглен граф. В това проучване основната идея е, че вместо броя на скрити слоеве, увеличава се броят на невроните на входа и скритите слоеве се разширяват по време на обучението на невронна мрежа. Удължаването на входния слой е свързано с факта, че всеки времеви

ред расте с поява на ново измерване. Целта на обучението е размерът на входния слой да бъде толкова голям, колкото размерът на пълната времева редица.

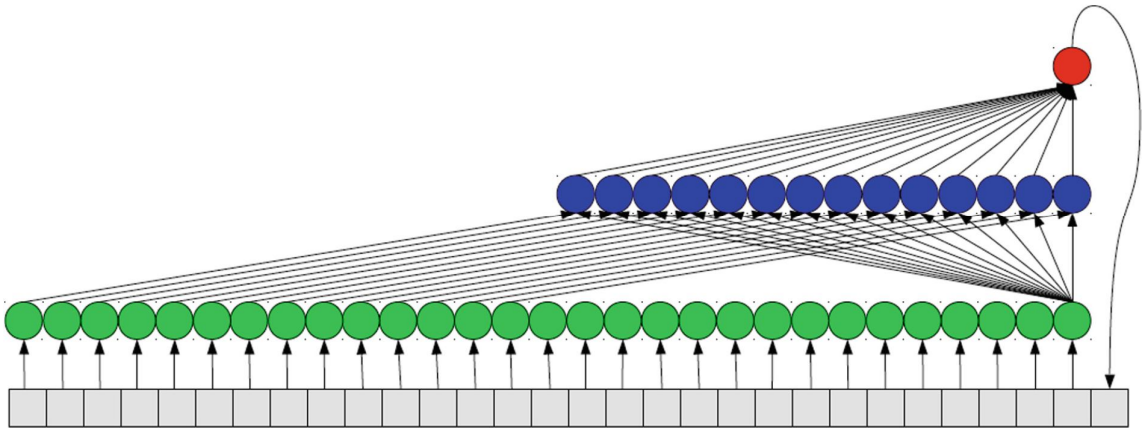
2.2.1 Предпоставки при моделирането

В предложения модел се използва набор от изкуствени невронни подмрежи и тези подмрежи се обединяват в обща изкуствена невронна мрежа. Най-малката изкуствена невронна подмрежа има 1-1-1 топология (фиг. 2.2.6 - вляво). Мрежата е обучена с примери, чийто вход има само една стойност. Целта в модела е прогноза за само една стойност напред във времето. Ето защо всички подмрежи имат само един изход. Всичките входни стойности се предоставят като примери за еластично обучение за обратно разпространение на грешката. Обучението спира на определено ниво на *epsilon* за пълна промяна на грешките на невронната мрежа.



Фиг. 2.2.6. Обучение на изкуствени невронни подмрежи с 1-1-1 топология (вляво) и 2-1-1 топология (вдясно).

След обучение на 1-1-1 топология стойностите на теглата на първата подмрежа се зареждат във втората подмрежа с 2-1-1 топология (Фиг. 2.2.6-вдясно). Трета подмрежа има 3-2-1 топология. Размерът на скрития слой се избира автоматично чрез алгоритъм за постепенно подрязване, внедрен в Encog Machine Learning Framework (http://www.heatonresearch.com/encog/). Топологиите на подмрежите се формират чрез добавяне на един неврон във входния слой и коригиране на размера на скрития слой с алгоритъм за инкрементално подрязване. Крайната цел е да се достигне n-m-1 топология (фиг. 2.8), която обхваща всички известни стойности на времеви редове.



Фиг. 2.2.8. Обучение на изкуствена невронна подмрежа с n-m-1 топология. Някои от връзките между входния и скрития слой не се визуализират за по-добър вид.

Общата идея зад предложения модел е постепенното обучение на състезателни по размер изкуствени невронни мрежи. Често срещаният проблем при обучението на изкуствени невронни мрежи е размерът на мрежата. Чрез разделянето на най-голямата мрежа в много по-малки мрежи се постига ускоряване на процеса на обучението. Предложеният модел има по-висока степен на самоадаптация, тъй като когато се появи нова стойност във времевия ред, размерът на изкуствената невронна мрежа нараства, което означава, че фазата на обучение и фазата на работа са едновременни.

2.2.2 Експерименти върху изследването

Експериментите се правят чрез JAVA програма, където изкуствените невронни мрежи се изпълняват чрез API, предоставен от Encog Machine Learning Framework. Като входни данни за експериментите се използват финансови времеви редове на FOREX пазар. Данните се вземат от ежедневна двумесечна търговия за валутни двойки EUR / USD и USD / JPY. Стойностите на времевите редове се мащабират в диапазона от -0,99 до +0,99 с правилото за мащабиране MinMax. Изходът на изкуствената невронна мрежа се пренастройва до първоначалния диапазон със същото правило, което се използва в обратна посока. Резултатите от експериментите все още са в диапазона на статистическата грешка, която идва от сложността на финансовите процеси и високочестотния шум вътре в данните.

Предложеният модел за самонадграждащи се трислойни MLP за прогнозиране на времеви редове е обещаващ подход за ускоряване на обучението на изкуствени невронни мрежи. Нарастващият размер на входния слой включва максимална информация, налична във времевите редове, но предложената процедура за обучение

на изкуствена невронна мрежа отчита, че по-старите стойности трябва да бъдат по-малко информативни.

2.3 Изводи

В тази глава се предлагат нови методи за анализ и прогнозиране на пазарни ценови движения чрез времеви редове и невронни мрежи.

В резултат на това са направени следните заключения:

1 В изследването до тук се обхващат основните аспекти от процеса по анализ – от дефиниране на проблема и поставяне на задачите, до представянето на методи за решаването им. Във всеки един от етапите се извършва представяне на реални доказателства, чрез които може да се идентифицира наличието на слабости или необходимост от намиране на по-рационален подход в разглежданата област.

2 Методите позволяват да се интегрират в системи за автоматизирана обработка и вземане на решения. Разработеният метод (MA Volatility Indicator) подобрява прецизността в осцилатор (Моментум) и работи в комбинацията от два инструмента EMA или SMA, като предлага нова методика за интерпретиране на резултатите при пазарни анализи и спомага за намаляване на риска от загуби и увеличаване на успех при автоматизираната търговия.

3 Предложеният алгоритъм за обучение чрез самонадграждане в трислойни MLP ускорява обучението на ANN при прогнозиране на финансови времеви редове.

Методите представени до тук, могат да бъдат прилагани от специалисти в различни области в системи с прогнозен характер, за вземане на решения, анализиращи събития и процеси базирани на времеви редове.

Глава 3. Решения за осигуряване на криптографска защита чрез приложение на времеви редове при криптографията и киберсигурността

В дисертацията се предлага подходът на времеви редове да се приложи към анализа на качеството на система за генериране на произволни числа (RNG) за осигуряване на криптографската защита в информационните системи. За текущото изследване от RNG се извлича числов масив, за да може да се анализират стойностите от случайните числа във времеви редове. Резултатите се изобразяват графично, където по-ясно стават видни произведените от генератора уязвими случайни числа.

3.1 Приложение на техники от времеви редове за анализ на генератор на произволни числа в областта на киберсигурността

RSA е асиметричен алгоритъм за криптиране, който позволява на всеки да изпраща криптирани съобщения, които само притежателя на частния ключ може да декодира. Принципът на работа може да се обясни накратко, като се генерира едно много голямо произволно число p, след това се генерира още едно такова число q и се изчислява тяхното произведение $x=p*q$, всъщност x е известен като публичен ключ.

3.1.1 Изследователите на (почти) секретния алгоритъм – слабости поради недостатъчна ентропия на RNG

На повърхността, RSA криптирането изглежда неуязвимо. Но според представеното изследване проблемът се крие в генераторите на случайни числа, които обезпечават алгоритъма. Уязвимостта е фундаментална и идва от там, че на RSA са необходими много големи числа, за да се създадат ключовете за криптиране, а генераторите в масовите компютърни системи са със значително по-малък капацитет. За това се налага да се използват псевдо генератор, който да се комбинира в качествените източници на ентропия, за да се изпълнят нуждите на алгоритъма. Чрез произлязла от генератора на числа началната стойност наречена семе (Seed), вложена в псевдо генератора и след трудоемки за компютърната система изчисления криптографските RSA ключове се генерират. Проблемът при устройства, като телефони, IoT, малки рутери и др. малки системи е още по-силно изразен, защото често те нямат достатъчно ресурс за тази трудоемка работа. За това в тях се залагат на готово предварително изчислени основни фактори необходими за съставянето на ключовете. Това значително ускорява процеса по генериране на RSA ключове при необходимост, но отваря голяма уязвимост в сигурността на криптографията. И отчитайки тези обезпокоителни наблюдения, те са достатъчно основание да се направи изследване по темата.

Съвременните критерии за надежден RSA ключ е минимум 2048 бита, като препоръчителната дължина е даже 4096 бита. При други изследвания също е установено, че между 4096, 8192 и 16384 бита RSA ключ, по-голямата сигурност на по-големите ключове е минимална. Причината също идва от ограниченията при генераторите на случайните числа. При по-големи RSA ключове са необходими

изключително големи истински случайни числа. Които в една компютърна система е крайно трудно да се получат.

Ако слабостите в криптографски функции не се осветяват, рискуваме да бъдат открити и да се използват от злонамерени лица без това да е известно на останалите. В заключение може да се каже, че слабостите не изхождат от грешка в аритметиката на RSA. Те идват от технологичната слабост, с която се прилага RSA.

3.2 Метод за оценка на уязвимостта на генераторите на случайни числа за криптографска защита в информационните системи

В предмета на изследването попада технологията на широко разпространения език за програмиране PHP. За нуждите на системи, разработвани с тази технология, за да се обезпечава необходимостта от случайни числа, PHP разполага със следните средства:

1. Линеен конгресен генератор (LCG), напр. `lcg_value()`
2. Алгоритъмът Marsenne-Twister, напр. `mt_rand()`
3. Локално поддържана функция C, т.е. `rand()`

Те се използват повторно и за функции като `array_rand()` и `uniqid()`, като недостатъкът на ентропията и генераторите на произволни числа на гореописаните функции се състои в лесното прогнозиране на бъдещите стойности на PRNG. Причината е, че първоначалните вътрешни състояния или SEED на PRNG са ограничени и изходът на стойности е в недостатъчен диапазон и това е предвидимо от лесно достъпните съвременни изчислителни ресурси. Често, за да получат стойност за SEED в PHP, разработчиците използват `mt_rand()` или следния скрипт, за да се използва автоматично:

```
<?php
mt_srand(3231153718);
for ($i=1; $i < 15; $i++) {
    echo mt_rand(), PHP_EOL;
}
```

Което поради слабата ентропия на предлаганите инструменти, води до риск възстановяването на SEED от нападател. Въпреки пасивния си характер, това всъщност е истинска уязвимост. За целта в изследването се създава симулация на истинска информационна система, което използва следния изходен код за генериране на токен за различните цели на приложението:

```
$newtoken = hash('sha512', mt_rand());
```

Генериране на токен по представения начин е хубав пример, като единично обръщане към `mt_rand()`, което се хешира с SHA512. Факт е, че в действителност, ако програмист приеме, че функциите на случайните стойности на PHP са "достатъчно случайни", той ще бъде много по-склонен да вгради прост модел на използване. Което е срещано многократно в практиката. Но използваният по-горе метод за генериране на маркери страда от един недостатък - случайните стойности са ограничени до цифри (т.е. неговата несигурност или ентропия е близка до незначителна). Ако се провери продукцията на `mt_getrandmax()`, ще се открие, че максималният произволен брой `mt_rand()` може да генерира само 2,147 милиарда. Този ограничен брой опции го прави уязвим за груба атака. При наличие на съвременна добра видео карта (GPU) и с помощта на специализиран софтуер за атака с груба сила като `hashcat`, такова изчисление може да се завърши само в рамките на няколко минути. Следователно използването на хеш за скриване на изхода на `mt_rand()` е безполезно.

За да се защити този тип система, трябва да се генерират случайни стойности с по-високо качество. За използване в нетривиални задачи, PHP изисква източници на ентропия от висок клас, които могат да бъдат осигурени от операционната система. В Linux обикновено се използва с `/dev/urandom`, освен ако не са инсталирани устройства с още по-висока ентропия. В Linux, с правилната настройка, редовен генератор на произволни числа, който е от типа PRNG (който е псевдо генератор на произволни числа), често се зарежда от източник на висока ентропия `/dev/random`, което го прави устойчив на атаки. Следователно всяка една софтуерна система разработвана с PHP, за да бъде добре защитена следва да се пренасочи към функциите за повторно използване на външната библиотека на OpenSSL. Като се извикват функциите `openssl_pseudo_random_bytes()` и `mcrypt_create_iv()`. Те са оптимизирани да използват криптографски защитен псевдослучайни генератор. Който е съобразен и интегриран с операционната система.

3.2.2. Разбиране на RNG Entropy в Linux

В операционната система Linux архитектурата за получаване на случайните числа има следния вид:

1. `/dev/random` е истински генератор на случайни числа, ако свърши ентропията блокира. Този инструмент си осигурява ентропията, събрана от

системните параметри по време на работата му като достъп до диск, мрежов трафик, състояние на паметта, преместване на мишката и други прекъсвания на системата;

2. `/dev/urandom` е генератор на псевдо произволни числа (PRNG) и той не се блокира поради изчерпването на ентропията. Може да се използва за рандомизиране на неограничен поток. Случайният поток се осигурява от PRNG структури и необходимите начални SEED стойности ще се презареждат периодично от /dev/random;
3. `/dev/hwrng` е допълнителен хардуер за истински случайни числа, който е специализиран и не е инсталиран в компютърните системи по подразбиране. Той осигурява шум от ентропия за поддържане на случайни числа;

Натрупаната ентропия в Linux система може да бъде проверена чрез следната команда:

```
$ cat /proc/sys/kernel/random/poolsize
4096
$ cat /proc/sys/kernel/random/entropy_avail
3868
```

където:

`/proc/sys/kernel/random/poolsize` се използва за деклариране на размера (в битове) на буфера Entropy Pool, например: Колко произволни числа трябва да съхраним, преди да спрем да „помпаме“ за повече.

`/proc/sys/kernel/random/entropy_avail` показва количеството (в битове) на текущо съхранени случайни числа в пула.

Чрез потребителската активност и работата на компютърната система, като мрежа, дискове, състояние на паметта, централен процесор, периферия и др. специалните функции в ядрото на Linux имат функции за непрекъснато набавяне на случайни числа. Което има за цел да компенсира непрестанната нужда от такива, при работа на компютърната система. Факт е, че колкото и да се опитва ядрото на операционната система да компенсира със случайни числа буферите, в определени моменти е възможно те да бъдат източвани много по-бързо. За нуждите на изследването лесно може да се предизвика такава ситуация, за да може да бъде наблюдаван този процес. Чрез следващата команда, просто се изхвърли всичко, което е в /dev/random генератора на произволни числа и се извежда на екрана:

```
$ hexdump /dev/random
0000000 d5c4 ff0a b8ef 9bdc ad95 480b e853 f0ef
0000010 e0cb 7c08 4bc4 daef 2b21 ea62 0eac 2c6c
0000020 d6bd 70e6 5d6f a7e3 0874 d52f 77df 6a2b
0000030 1909 efe8 9964 acee 2aad 2522 4ddb 1d0b
```

В същия момент може в паралелно отворен команден терминал да се изведе състоянието на буфера на ентропия, като съдържанието се обновява всяка секунда. За целта е необходимо да се стартира следната комбинация от команди:

```
$ watch -n 1 cat /proc/sys/kernel/random/entropy_avail
```

Като резултат наличието на ентропия ще започне да спада, като неговото състояние ще стигне критични стойности, дори и до нула. С натисне на `Ctrl-C` се спира това безсмислено разхищение. Може би никога не трябва да се прави това на практика, особено на реална сървърна система, освен с изследователска цел разбира се. Но често системите имат проблеми с натрупването на ентропия в буфера и резултатът изглежда смущаващ:

```
$ cat /proc/sys/kernel/random/entropy_avail
96
```

От представения пример машината произведе ентропиен резултат от 96 бита и увеличаването на тази стойност е твърде бавно и недостатъчно. Причините за това могат да са разнородни. Например от липса на специфичен хардуер, неправилни настройки, виртуализация, твърде голяма активност със случайните числа на системата и невъзможност да се компенсира консумацията на случайни стойности и др. Едно възможно решения е да се стартира специализиран софтуер подпомагащ събирането на случайни числа. Това е демон, който е проектиран да използва всякакви събития, които могат да се считат за сравнително случайни при работата на машината, за да се произведат повече и по-качествени случайни числа. Например процесорното „трептене“, промяната в състоянието на паметта, входно изходни операции, мрежов трафик могат да добавят още ентропия към буфера на системата. Инсталирането на това решение и основната настройка в системата са следните:

```
# apt install haveged
# systemctl start rngd
# update-rc.d haveged defaults
```

```
# rngd -r /dev/urandom
```

На система със сравнително умерен трафик:

```
# pv /dev/random > /dev/null
  40 B 0:00:15 [   0 B/s] [                    <=>              ]
  52 B 0:00:23 [   0 B/s] [                           <=>       ]
  58 B 0:00:25 [5.81 B/s] [                              <=>    ]
  64 B 0:00:30 [6.05 B/s] [                                 <=> ]
^C
# systemctl start haveged
# pv /dev/random > /dev/null
7.12MiB 0:00:05 [1.43MiB/s] [               <=>               ]
15.7MiB 0:00:11 [1.44MiB/s] [                       <=>         ]
27.2MiB 0:00:19 [1.46MiB/s] [                           <=>     ]
  43MiB 0:00:30 [1.47MiB/s] [                                <=> ]
^C
```

С помощта на командата `pv` може да се види колко данни се предават за целта. От показания поток на данните се вижда, че преди `haveged` се получаваха 2.1 бита в секунда (B / s), докато след това се получават ~ 1.5 MB / sec.

3.2.3. Времеви редове за генератори на случайни числа

Спецификата на RNG и PRNG позволява те да бъдат анализирани чрез техники за анализ и прогнозиране на времеви редове, тъй като улавянето на потока на изходните числови стойности е последователност и само по себе си е подредена последователно във времето. Такъв поток от числови стойности може да бъде описан, както следва:

$$N = T * V$$

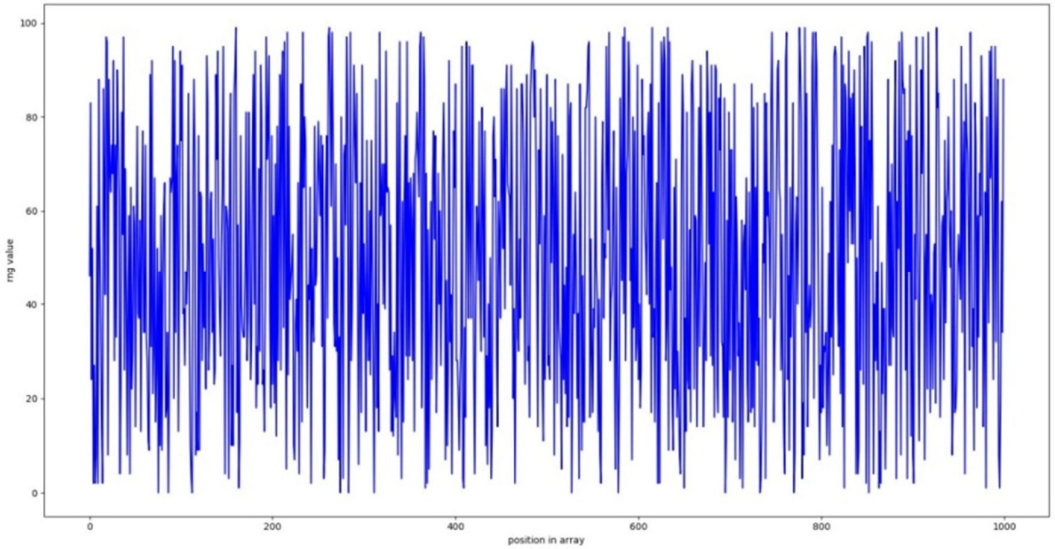
където: N - дължината на числовия ред, T - време (продължителност) на генерирането на числа, V - брой генерирани числа за единица време.

Така, че чрез времевите редове е възможно да се определи качеството на ентропията във времето. Ако даден генератор на случайни числа не е много надежден, то неговите слабости биха могли да се намерят за по-кратък времеви ред с данни, за които ще са необходими по-малко ресурси за обработка и анализ. За нуждите на текущото изследване ще се използва числов масив, който няма да бъде създаден от висококачествен генератор на случайни числа, а от посредствен такъв. Идеята е да се

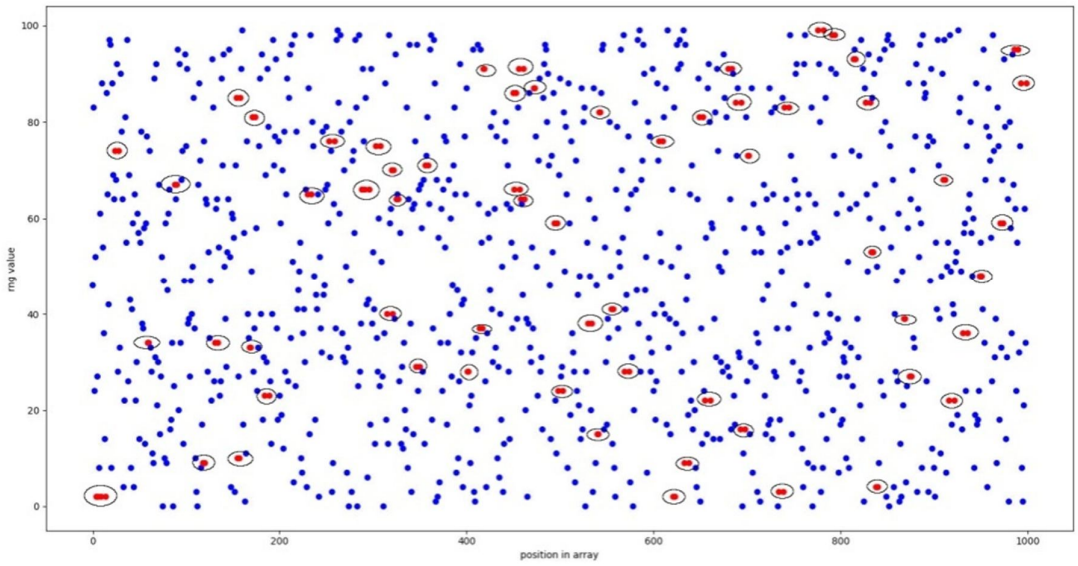
приложи подходът и да се анализира времевата линия от случайни стойности от среден клас компютърна система, с каквато най-често всеки разполага.

3.2.4. Проучване на генератори на случайни числа с времеви редове

Времевите редове като стохастичен процес могат да бъдат използвани за анализ и на RNG/PRNG. За тази цел е разработен алгоритъм за откриване на повтарящи се модели (patterns) от данни в генерираните от RNG времеви редове. За да се съберат данни от генераторите на случайни числа във времеви ред за нуждите на изследването се използва специално написана програма. Посредством отново написана за целта програма, събраните данни със случайни числа се представят графично, което помага за по-лесно забелязване на важните елементи от времевите редове (фиг.3.2 и фиг.3.3). На пръв поглед с резултатите от данните от System.Random на фиг. 3.2 всичко е наред и е възможно да се мисли, че имат добро качество на ентропия. Но нека предложа още един начин в друг графичен изглед, за да се уверим в преценката си.



Фиг. 3.2. Визуализиране на данните получени от System.Random, като линия на шум.



Фиг. 3.3. Представяне на данните от System.Random във визуализация тип поле с точки

Представянето на едни и същи данни с различна графична интерпретация, може да помогне за разкриването на някои проблеми с качеството на изследваните стойности. На фиг.3.2 и фиг.3.3 са изобразени графики на едни и същи данни. Като при фиг.3.2. качеството на случайните числа според графиката може да се приема за добро. Но след преглед на графиката от фиг.3.3 става ясно, че има чести случай на повторения, които програмата прихваща, като модел на повторяемост. Съответните стойности се разпознават и оцветяват с червено и са оградени в кръг за по-добра видимост.

Визуализацията на фиг. 3.3 показва слабостите на обработените резултати. Моделите на повторното появяване се срещат периодично във времето. Тези случаи са оцветени в червено от разработената програма, като се използват предварително определени модели за прогнозиране, както беше споменато по-рано. В конкретния случай на разглежданите времеви редове, тези случаи са 65 случая от масив с 1000 стойности. Може да се каже, че 6,5% прогнозни числа от генерирания RNG масив са значителен резултат. Обикновено последователност от псевдослучайни числа се инициира от SEED вътре в PRNG (Koeune, 2005). Ако такъв генератор на произволни числа се използва в криптографията, произведените от него SEED стойности могат да бъдат атакувани успешно. Чрез прогнозиране на следваща стойност SEED или чрез наблюдение на предадени криптирани данни, стойностите в основата на системата за криптиране могат да бъдат възприети в определен момент.

3.3 Пренебрегнати рискове от киберсигурност в доставчиците на услуги за публичен Интернет хостинг

До тук изследването за анализи на качеството на RNG и областите, в които се срещат проблеми, успя да засегне криптографските алгоритми, езици за програмиране и операционни системи. Сега фокусът се измества върху масово предлагани публични хостинг услуги. В тази дисертация за настоящото изследване е използван уеб хостинг доставчик, който е един от популярните в бранша. Услугата за уеб приложения е инсталирана на масово предлаган нает споделен хостинг. Добавен е уеб сертификат и SSL достъпът е активиран, като всичко работи на стандартните портове за комуникация. На първа линия на защита между клиент и сървър излизат критографските шифри, поддържани от хостинг сървъра. Ако те са актуални и между тях не се срещат уязвими и вече остарели и компрометирани във времето такива, може да се счита, че протоколът за комуникация е достатъчно добре подсигурен.

Извършен бе тест, като са сканирани криптографските протоколи, обезпечаващи връзката между клиент и сървър (хостинг услугата). Установено е, че от списъка с криптографски протоколи, които сървърът предлага участват: TLSv1.0 и TLSv1.1, които изобщо не трябва да се поддържат и предлагат, тъй като имат отдавна установени слабости и не трябва да се използват. Друг протокол, който сървърът поддържа е TLSv1.2, който е все още актуален и е одобрен за използване, но не в пълния му вид. В него се съдържат криптографски шифри, които трябва да бъдат извадени, но сървърът ги предлага за комуникация, което също е съществена уязвимост в сигурността на предоставяната услуга. Анализът на протоколите и шифрите също така установи още един съществен недостатък. Протоколът TLSv1.3 не се поддържа изобщо, това за момента е най-актуалния и сигурен протокол от семейството на TLS за тунелна свързаност.

След проверката на криптографските протоколи и шифри, поддържани за комуникация, изследването се премести върху по-чувствителната тема – генераторите на случайни числа. Понеже дори и при най-актуалните протоколи и шифри за защита, ако случайните числа не са достатъчно качествени и случайни, рискът да падне цялата криптираща защита е много голям. За да се извърши този анализ е създадена компютърна програма, която установява свързаност до сървъра по наличните криптографски протоколи за защита между клиент сървър. В конкретния случай е използван TLSv1.2 и във фазата на установяване на свързаност, програмата взима

генерираните случайни числа от сървъра и ги записва във файл, като времеви ред. Въпросната програма се изпълнява в цикъл, докато събере достатъчно количество данни за анализ.

Събраните данни от случайни числа са подложени на анализ чрез специализираният софтуер с отворен код за анализ на случайни числа използвани в криптографията Dieharder на Робърт Г. Браун (Brown, 2021). Изпълнени са симулация на 114 теста, както и проверка на качеството на числата и по стандарта за киберсигурност на генератори на случайни числа FIPS-140. Обобщено данните от теста за симулация на случайни числа са:

• Само 25 теста са преминали успешно;

• Неуспешни, които имат компрометирана /предсказуема/ стойност и следователно откриваема криптография са 76;

• Уязвими, където криптографията може да бъде разкрита с относително добър компютърен хардуер са 13;

От представените резултати може да се направи заключение, че заради слабостите в случайните числа и при установеното нарушаване на криптографската защита, рискът за успех при кибератаки за компрометиране на криптографията е критично висок. Причините за това може да са разнородни, но най-често срещаната от тях е, че хостинг доставчиците често поемат повече клиенти с техните приложения, от колкото капацитета на киберзащитата на сървърите им може да поеме. Работата на много приложения и клиенти едновременно, непрестанно източват криптографията и случайните числа на сървъра.

Решенията от страна на клиента, които се допускат в случая, е да се използва частен хостинг върху собствена инфраструктура, където няма да се допусне прекомерното натоварване от описания вид. При невъзможност да се осигури обаче непрекъсваемост за хардуерна конфигурация и подходящо място, като сървърно помещение. По-добре е да се наеме VPS сървър, който ще е само под контрола на един клиент и също проблемът ще се избегне. От страна на хостинг доставчика обаче, също може да се предприемат действия за повишаване капацитета на киберзащитата. Следва да се приложат техниките за конфигуриране на правилното функциониране и повишаване капацитета на ентропия в Linux, описани в раздел 3.2.2 „Разбиране на RNG Entropy в Linux“ на тази дисертация.

След правилната настройка на системата, може да се прибегне до друг нетрадиционен подход, познавайки принципа на работа на събиране на ентропия в

буферите си от операционната система Linux. Може да се напише програма, която да генерира редици от събития, които няма да затормозят особено системата, но ще създадат множество процеси подпомагащи събирането на ентропия:

```
#!/bin/sh
## list of sites using round-robin DNS
ROUND_ROBINS="www.yahoo.com google.com twitter.com outlook.com"
## Entropy stard and end value limits
STOP_LIMIT="3800"
START_LIMIT="3000"


until [ "$(cat /proc/sys/kernel/random/entropy_avail)" -gt
"$STOP_LIMIT" ]
    do while [ "$(cat /proc/sys/kernel/random/entropy_avail)" -lt
"$START_LIMIT" ]
        do for thing in "/tmp/loyeyoung" "/tmp/sueellen"
"/tmp/rootdev" "/tmp/files"
            do echo $thing ====================================
                touch /tmp/toss
                for robins in $ROUND_ROBINS
                    do nslookup "$robins" 8.8.8.8 > /tmp/toss
                        nslookup “$robins” 9.9.9.9 >> /tmp/toss
                        nslookup “$robins” 192.168.2.3 >> /tmp/toss
                        nslookup “$robins” >> /tmp/toss
                        cat /tmp/toss
                        mkdir $thing -p
                        cp /tmp/toss $thing/toss
                        cat $thing/toss
                        rm -f /tmp/toss
                        rm -f $thing/toss
                    done
                done
            done
        done
```

Представения програмен скрипт е съвсем базов и би могъл да бъде надграден и съставян и на други програмни или скриптови езици. Въпреки семплия вид, успява да даде очакваните резултати и покрие нуждите на текущото изследване. Скоростта на натрупване на ентропия се подобри. Което допринася въпросната система да понася по-големи натоварвания върху генерирането RNG стойности. Начинът на действие е както е зададен в момента е, че изпълнението на допълнителните операциите в памет, процесор, диск и мрежа, ще се активират при достигане на стойност в буфера за ентропия под 3000. Също така, би могло предоставеното решение да се използва в комбинация с хардуерни решения, подпомагащи криптографските алгоритми и ентропията на случайните числа, което и компанията Intel предлага при своите процесори.

Наименованието на модула за подпомагане генерирането на случайни числа е Intel Secure Key, предишното му кодово име е Bull Mountain Technology. С това името Intel определя в процесорите си разширението за архитектура Intel64 и IA-32 RDRAND и свързаната с него хардуерна реализация на Digital Random Number Generator (DRNG). Освен всичко друго, DRNG, използвайки инструкцията RDRAND може да е изключително полезен при генериране на висококачествени ключове за криптографски протоколи. Следователно трябва да се провери, дали текущата система разполага с такива процесори и би могло нейната конфигурация да бъде обновена. При наличие на компютърна система с Linux операционна система, проверката може да стане освен чрез техническата документация на чиповете от производителя и чрез следната комбинация от команди:

```
$ cat /proc/cpuinfo | grep -i rdrand | echo $?
0
```

Като резултат 0 означава, че е наличен флаг RDRAND и процесорът може да бъде включен за подобряване на криптографските функции на системата по следния начин:

```
# apt install rng-tools-debian
# /etc/init.d/rng-tools-debian start
# /etc/init.d/rng-tools-debian status
* rng-tools-debian.service - LSB: rng-tools (Debian variant)
   Loaded: loaded (/etc/init.d/rng-tools-debian; generated)
```

```
   Active: active (running) since Fri 2020-11-28 17:30:54 EET; 3min
10s ago
     Docs: man:systemd-sysv-generator(8)
    Tasks: 4 (limit: 4915)
   Memory: 1.3M
   CGroup: /system.slice/rng-tools-debian.service
           `-3597 /usr/sbin/rngd -r /dev/hwrng
$ cat /proc/sys/kernel/random/entropy_avail
4096
```

Резултатите показват, че скоростта на събиране на ентропия за нашият случай надхвърля скоростта на нейното консумиране.

3.4 Резултати в реална технологична инфраструктура

Предложения подход за подобряване на киберсигурността в криптографията и генераторите на случайни числа при натоварени сървърни системи с публични услуги е приложен в технологичната инфраструктурата на института ИИКТ-БАН. Използваната хардуерна конфигурация е от среден клас, като е съобразена със сложността на изпълняваната задача. Сървърът е оборудван с един шест ядрен процесор Xeon(R) E-2236 от второ поколение и версия 6, 32GB RAM и два твърди диска в конфигурация с RAID1. Оперативния сървър с публичните услуги функционират върху Linux и всичките услуги са изцяло и от софтуер с отворен код. Функционират върху виртуална машина, като физическата машина е само виртуален хост, което е еквивалентно със ситуацията с разглежданите масови услуги, които са в предмета на текущото изследване за киберустойчивостта на криптографската защита. Сървърните услуги, изпълнявани от виртуалната машина са:

- мейл сървър, към момента с 242 потребителски акаунта. Достъпен чрез SMTP, POP3, IMAP, като всички те са защитени с криптографски комуникационен протокол TLSv1.2 и TLSv1.3. Удостоверяват се със сървърен сертификат за установяване на TLS сесии с асиметричен алгоритъм от типа елиптична крива secp384r1. Свързването до услугата не може да се осъществи без криптиране на комуникацията;
- Уеб мейл, който позволява на всичките 242 потребителя да оперират с пощата си и през уеб браузър. Комуникацията е защитена чрез криптографския комуникационен протокол TLSv1.2 и TLSv1.3. Удостоверяват се със сървърен

сертификат за установяване на TLS сесии с асиметричен алгоритъм от типа елиптична крива secp384r1. Свързването до услугата не може да се осъществи без криптиране на комуникацията;

- Уеб портал на Институтът по информационни и комуникационни технологии към Българската академия на науките, което е основното уеб пространство на института. Съдържа информация за дейността, два научни журнала, както и структурна информация. Комуникацията е защитена чрез криптографски комуникационен протокол TLSv1.2 и TLSv1.3 и сървърен сертификат за установяване на сесии със асиметричен алгоритъм с елиптична крива secp384r1. Удостоверяват се със сървърен сертификат за установяване на TLS сесии с асиметричен алгоритъм от типа елиптична крива secp384r1. Не се позволява свързване до услугата по не криптиран канал;
- Услуга за отдалечена администрация SSH с най-високата степен на криптографска защита, предлагана от протокола към момента. Идентификацията на потребител по SSH е само чрез криптографски ключове, не се допускат пароли;
- Услуга за отдалечено управление на Уеб съдържанието FTP. Комуникацията е защитена чрез криптографски комуникационен протокол TLSv1.2 и TLSv1.3 и сървърен сертификат за установяване на сесии със асиметричен алгоритъм с елиптична крива secp384r1. Удостоверяват се със сървърен сертификат за установяване на TLS сесии с асиметричен алгоритъм от типа елиптична крива secp384r1. Не се позволява свързване до услугата по не криптиран канал;

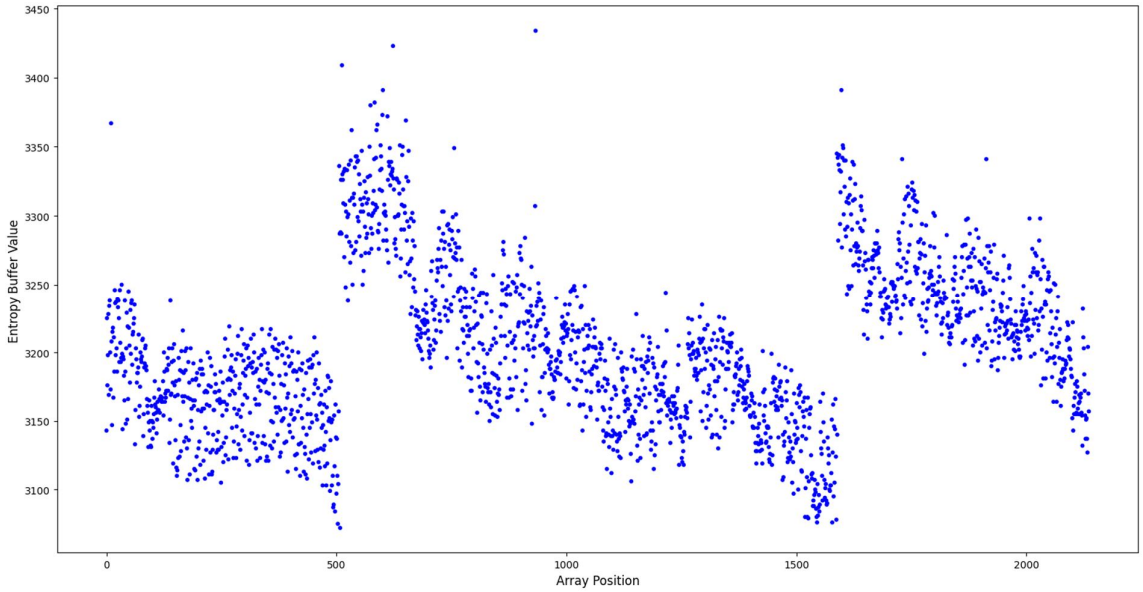
За всички услуги приоритетен протокол за криптирана свързаност е най-новият и сигурен протокол TLSv1.3, но ако се окаже, че клиента не го поддържа се минава на протокол TLSv1.2. Последният е оставен само за съвместимост, като от него са премахнати всички криптографски алгоритми в които са открити уязвимости.

Установено е, че към настоящия момент сървърът се ползва значително интензивно от потребителите на ИИКТ и външни потребители на Интернет. Нивото на криптографска защита е на най-високото към момента според установените стандарти и не са правени никакви компромиси с криптографските протоколи или шифри. Като доказателство за качеството на киберзащитата с криптографски средства е приложен резултат от тест чрез скенер на SSL Labs за нивото на криптиране при предлаганите TLS протоколи върху наличните услуги. Резултатите от този тест са извлечени на базата на актуалните изискванията за криптографска защита към момента, които са

утвърдени от международните лаборатории по криптографска защита FIPS и NIST за САЩ, и Common criteria за Европа. От резултатите е видно, че протоколите и средствата за криптиране са от най-актуалните към сегашния момент. Оценката на всички тестове е най-високата възможна A+. Нивото на защита на HTTP протокола, който комуникира с браузъра, чрез TLS тунелът също е с максималното ниво на защита A+.

Извършени са тестове и на качеството на ентропия, чрез методите представени в дисертацията. От командния shell на сървъра са приложени два утвърдени метода, първият проверява качеството на ентропия по FIPS с rngtest, а вторият с инструмента за анализ dieharder. Всички тестове на ентропията на случайните числа издържат с най-високия възможен резултат според критериите на въпросните програми.

Въпреки добрите резултати е направено още едно изследване, според което става ясно, дали в моменти на висока потребителската активност и интензивното натоварване на криптографията ще доведат до изчерпване капацитета на случайните числа. За целта е съставен програмен скрипт, който на всеки 10 минути записва размерът на натрупаната в буфера ентропия. Тази статистика се събира в рамките на половин месец. За да стане ясно дали има моменти, които водят до изчерпване на буфера с ентропия по-бързо от колкото е неговото зареждане от системата. Като резултат след половин месец събиране на данни във времеви ред, се получиха 2137 стойности. Тези стойности са изведени в графичен вид и са изобразени на фиг.3.10:



Фиг.3.10 Ниво на ентропия в различни моменти от времето

От фиг.3.10 личи, че системата е имала пикове на по-интензивна дейност, които са водели до силно черпене от натрупаната в буфера ентропия. За това стойностите на моменти рязко падат. Предложеният в изследването подход върху системата обаче, успява да компенсира високата консумация на ентропия. Макар на графиката амплитудите между максимални и минимални стойности да изглеждат широки, то стойностите са в тесен диапазон, с ниво на ентропия между 3000 и малко под 3450. Липсата на стойности под 3000 показва, че системата се намира в много добро здраве и дори е способна да поеме по-големи натоварвания, защото стойностите са далеч от критичните. Като се вземат предвид всички тези резултати от реалната работна среда на сървъра, е на лице доказателството за ефективността на предлагания подход. Следователно може да се счита, че предлаганият подход може да бъде от полза и да подпомага различните Интернет системи и решения.

3.4 Изводи

Изследването на представените услуги и тяхното ниво на киберсигурност е от ключово значение за по-сигурен преход към съвременната дигитална трансформация. Бързото прехвърляне на всички социални и икономически дейности към дигитални платформи доказва, че съществуващата технологична инфраструктура може да отговори на днешните предизвикателства за дигитална трансформация. Ползите от това в икономическо и екологично отношение са неоспорими. Но по отношение на киберсигурността, много от настоящите ИТ услуги все още изостават. Увеличаването на степента на успех на киберпрестъпленията може да доведе до загуба на доверието в технологиите и възпрепятстването на тези процеси, което ще засегне и научно-техническия прогрес. Също така ще повлияе на забавянето в развитието и на много други свързани области в икономиката, сигурността, технологиите.

Прилагането на математически и статистически анализи с времеви редове за решаване на проблеми в киберсигурността е ефективно. Предлаганите тук подходи, може да се комбинират и с други техники и методи за анализ на киберсигурността, за да са по-комплексни и ефективни. В тази дисертационна работа е разработен метод за изследване на качество на RNG и PRNG в информационна система чрез прилагане на времеви редове. Методът позволява да се повиши качеството на ентропия при използването на криптография, осигуряваща различни Интернет услуги. Разработен е алгоритъмът за откриване на повтарящи се модели (patterns) от данни в генерираните от RNG времеви редове. Проведено е изследване на криптографски тестове и качеството

на ентропия върху работещи в реални условия натоварени сървърни системи с публични Интернет услуги.

Глава4. Софтуерни подходи при работа с големи масиви от данни и ограничени компютърни ресурси с език за програмиране R

4.1 Програмният език R

Програмният език R е продукт, разполагащ с мощни инструменти за статистически изчисления и анализи. R едновременно е програмен език и софтуерна среда (Borcard, 2011), (The R, 2017). Компилира се и работи на различни операционни системи, като UNIX платформи, Linux, Windows и MacOS. Езикът R предлага широко разнообразие от статистически техники като, например линейно и нелинейно моделиране, класически статистически тестове, анализ на времеви редове, класификация, групиране и др., както и графични техники и е изключително разширяем (Long, 2015).

Въпреки многото предимства на R, богатството на неговите статистически модели и инструменти за обработка на данни, както и мощните способности за визуализация, изникват проблеми при работа с големи обеми от данни. Ограниченията на R произлизат от това, че той е проектиран да оперира в режим на изчисления само в единичен процес (на единично ядро на процесор) и при данните, заредени наведнъж в оперативната памет.

4.2 Преодоляване на проблеми на работа с големите данни чрез използване на микропроцесор с много ядра

Паралелното програмно изчисляване на повече от едно ядро на процесор е възможно чрез прекомпилиране и добавяне на някои програмни компоненти в R. Това е възможно, поради факта, че R е система с отворен код и това е едно от предимствата, което носи тази концепция.

4.3 Методи за оптимизиране на обеми от данни

Една от всеизвестните особености на езика R е, че зарежда всички данни с които оперира в RAM паметта на компютърната система, което при работа с големи данни би било критично, дори и на мощни системи с голям ресурс. Като начин за решаване на този проблем, в дисертацията се разглеждат начини за зареждане на данните в паметта,

като се изключват данните с некоректно съдържание още в момента на тяхното зареждане.

В някои статистически изследвания не е необходимо да се зареждат всичките данни, а само определени времеви рамки, за да се направи приблизителен статистически анализ в отрязък от време. В такъв случай, може да се приложат методи за позиционирано прочитане, предложени в дисертационния труд. Така става възможно да се обработи само определен отрязък от данните, разположени във файл с големите данни. Друг често възникващ проблем е, при зареждане на големи данни е, че след зареждане в паметта и обработка, някои от данните вече не са необходими, но продължават да заемат значителни обеми памет. В дисертацията е представен начин за редуциране на данните в паметта, като се премахват излишните от тях и се освобождава памет.

4.4 Изводи

Приносът на автора е, че чрез този материал със средствата на език за програмиране R се подпомага решаването на проблеми при работа с големи масиви от данни при ограничени компютърни ресурси. В тази дисертация са разработени софтуерни техники за оптимизиране на компютърната памет при работата с големи данни.

В заключение може да се каже, че с представените до тук примери не може да се изчерпа темата за оптимизираното зареждане на данни при работа с език за програмиране R. Работа с реални данни винаги е предизвикателство (Baumer, 2017). Но представените до тук техники са между добрите практики и са често използвани, те биха могли да се комбинират и с други подходи за решаването на проблеми в тази област.

Заключение - резюме на получените резултати

В дисертационния труд подробно са изследвани методи и средства за използване на времеви редове при решаване на различни задачи, възникващи в съвременните приложения на информационни технологии и системи.

Предложен е метод озаглавен MA Volatility Indicator за подобряване прецизността в осцилатор (Моментум). MA Volatility Indicator работи в комбинацията от два инструмента EMA или SMA и предлага нова методика за интерпретиране на резултатите, което допринася за откриване на нива за свръх покупка и свръх продажба

при пазарната тенденция. Всички използвани в изследването инструменти EMA, SMA и Моментум, както и MA Volatility Indicator използват времеви редове.

Разгледана е приложимостта на апарата на невронните мрежи за прогнозиране на времеви редове във финансовата област. Показано е, че с нов модел на представяне на входните данни, характерни за финансови показатели, се получава по-висока степен на самоадаптация при обучение на невронната мрежа. Проведените експерименти потвърждават сложността на финансовите процеси и наличието на високочестотен шум в данните.

Разработен е метод за изследване на качество на RNG и PRNG в информационна система чрез прилагане на времеви редове за да се повиши ентропия на при използването на криптография, осигуряваща различни Интернет услуги. По този начин се допринася за по-добрата киберсигурност на ИТ инфраструктура за цифрови ресурси и защитата на данни. В дисертационната работа темата за криптографията получи специално внимание, поради нейното критичното значение. При пропускането само на един риск в киберсигурността е възможно да бъдат компрометирани всички ИТ услуги.

Практическите резултати от реалния експеримент показаха, че е намерено златното съотношение между масови услуги и действителните изисквания за киберсигурност.

С оглед на работата, извършена в този дисертационен труд и резултатите, получени в хода на изследванията и изложени по-горе, могат да бъдат формулирани следните научно-приложни резултати:

1. Разработен е метод, озаглавен MA Volatility Indicator, за комбиниране на индикатори за откриване на ценови движения с нови подходи при използване на времеви редове от финансовите данни.

2. Приложен е апаратът на изкуствени невронни мрежи с цел изследване на финансови времеви редове. Разработен е алгоритъм за обучение на невронната мрежа чрез увеличаване на размера на входа на невронна мрежа и създаване на хибридна структура, като е предложен модел за самонадграждащи се трислойни MLP.

3. Разработен е метод за повишаване на криптографската защита в информационните системи на базата на изследвания на качеството на генераторите на произволни числа.

4. Проведени са експериментални изследвания за решаване на проблемите с киберсигурността в публични широко разпространени хостинг услуги. Получените

резултати потвърждават валидността на предложения метод за повишаване на киберсигурността.

5. Разработени са програмни методи за ефективна работа с големи данни със средства на езика R.

6. Разработените методи за повишаване на криптографска защита са имплементирани в технологичната инфраструктурата на ИИКТ-БАН. Проведено е изследване на криптографски тестове и качеството на ентропия върху работещи в реални условия натоварени сървърни системи с публични Интернет услуги.

Насоки за бъдещи изследвания

Насоките за бъдещи изследвания по тематиката на дисертацията включват:

- Имплементиране на методът MA Volatility Indicator и прилагането му в комбинация и с други методи за анализ и прогнозиране на пазарни ценови тенденции;
- Прилагане на методът MA Volatility Indicator към автоматизирани системи за анализ на пазарни тенденции и извличане на сигнали за взимане на решения;
- Провеждане на още изследвания в областта на обучаващи алгоритми и системи с невронни мрежи за анализ и прогнозиране на времеви редове;
- Развиването на нови методи за увеличаване на криптографската защита в информационните системи;
- Изследване на комбинация на разработен метод с други методи и системи за анализ на RNG в криптография и други технологични области, което да съдейства за създаване и усъвършенстване на RNG, както и за по-точно определяне на спектъра от задачи, които генераторът може да изпълнява добре;
- Намиране на още подходи за зареждане и филтриране на големите данни с цел по-ефективната им обработка.

Публикации по темата на дисертационния труд

1 **Иван Благоев**, Николай Докев, Комбиниране на Моментум с един метод за прогнозиране на пазарни ценови движения за по-точни резултати (Combination of Momentum with One Method for Forecasting of Market Trends to Improve the Results), Международна научна конференция “УНИТЕХ’17” – Габрово, 2017 Selected papers, ISSN 2603-378X, pp. II-265-II-270

2 **Ivan Blagoev**, Методи за оптимизирано използване на компютърна памет при зареждане на данни със средствата на език за програмиране R (Methods for

Optimized Use of Computer Memory during Data Loads with R Programming Languages), International Conference “Automatics and Informatics’2017”, 4-6 October 2017, Sofia, Bulgaria, ISSN:1313-1850, pp.213-215.

3 **Blagoev I.,** Improving the Momentum Oscillator Accuracy by a Method for Forecasting of Market Price Movements, Сборник доклади от международна конференция, НВУ "Васил Левски", 14-15 юни 2018, Том 9, стр. 177-185. (ceeol.com)

4 **Blagoev I.,** Method for more reliable users’ authentication in internet, Сборник доклади от международна конференция, НВУ "Васил Левски", 14-15 юни 2018, Том 9, стр. 167-176. (ceeol.com)

5 **Blagoev, I.,** Using R Programming Language for Processing of Large Data Sets, Proc. Int. Conf. Big Data, Knowledge and Control Systems Engineering – BdKCSE’2018, 21-22 November 2018, Sofia, Bulgaria ISSN 2367-6450, pp. 91-98.

6 **Ivan Blagoev**, Application of Time Series Techniques for Random Number Generator Analysis, Proceedings of XXII Int. Conference DCCN 2019, September 23-27, 2019, Moscow, Russia, pp.437-446. ISBN 978-5-209-09683-2, 2019 (РИНЦ).

7 **Blagoev I.,** Neglected Cybersecurity Risks in the Public Internet Hosting Service Providers. *Information&Security International Journal* - ISIJ, 47, no. 1, pp. 62-76 (2020)

8 Balabanov T.D., **Blagoev I.I.,** Dineva K.I. (2018) Self Rising Tri Layers MLP for Time Series Forecasting. In: Vishnevskiy V., Kozyrev D. (eds) Distributed Computer and Communication Networks. DCCN 2018. Communications in Computer and Information Science, vol 919. Springer, Cham. https://doi.org/10.1007/978-3-319-99447-5_50, pp. 577-584, **SJR:0.188**

9 **Blagoev I.** (2020) Method for Evaluating the Vulnerability of Random Number Generators for Cryptographic Protection in Information Systems. In: Dimov I., Fidanova S. (eds) Advances in High Performance Computing. HPC 2019. Studies in Computational Intelligence, vol 902. Springer, Cham. https://doi.org/10.1007/978-3-030-55347-0_33 **SJR:0.215**

Забелязани цитирания

I **Blagoev, I.,** 2018. Using R Programming Language for Processing of Large Data Sets, Proc. Int. Conf. Big Data, Knowledge and Control Systems Engineering – BdKCSE’2018, pp. 91-98.

Цитира се в:

1 Dineva, K., Atanasova, T.: Regression Analysis on Data Received from Modular IoT System. ESM'2019, EUROSIS-ETI, ISBN: 978-9492859-09-9, EAN: 9789492859099, pp.114-118, 2019

2 Ivaylo Blagoev, G. Vassileva and V. Monov, "Methodology for content preparation of online courses," 2020 International Conference Automatics and Informatics (ICAI), Varna, Bulgaria, 2020, pp. 1-4, doi: 10.1109/ICAI50593.2020.9311364.

II **Blagoev I**., Neglected Cybersecurity Risks in the Public Internet Hosting Service Providers. *Information&Security International Journal* - ISIJ, 47, no. 1, pp. 62-76 (2020)

Цитира се в:

3 M Terzieva, D Karastoyanov, ICT for Innovation in Advanced Banking, PROBLEMS OF ENGINEERING CYBERNETICS AND ROBOTICS • 2020 • Vol. 73, pp. 47-54 p-ISSN: 2738-7356; e-ISSN: 2738-7364, doi: 10.7546/PECR.73.20.05

III **Blagoev I.**, Method for more reliable users' authentication in internet, Сборник доклади от международна конференция, НВУ "Васил Левски", 14-15 юни 2018, Том 9, стр. 167-176.

Цитира се в:

4 Ivaylo Blagoev, Gergana Vassileva and Vladimir Monov, "Methodology for content preparation of online courses," 2020 International Conference Automatics and Informatics (ICAI), IEEE, Varna, Bulgaria, 2020, pp. 1-4, doi: 10.1109/ICAI50593.2020.9311364.

5 Dineva, K., Atanasova, T.: Security in IoT Systems. Proceedings 19th International Multidisciplinary Scientific Geoconference SGEM 2019, 19, 2.1, ISBN:978-619-7408-79-9, ISSN:1314-2704, DOI:10.5593/sgem2019/2.1, 576-577. SJR (Scopus):0.232 Q4

Участие в проекти

1 Национална научна програма „Информационни и комуникационни технологии за единен цифров пазар в науката, образованието и сигурността“ (ИКТ в НОС) - 2018-2021.

2 Проект Зора по Заповед Но 147/14.06.2019 "Цифров и кибер устойчив ИИКТ"

Награди

1. Награда на ИИКТ-БАН за отлични научни постижения през 2019 г. в категория „Докторанти“.

Библиография

1 Atanasova, T., Barova, M.: Exploratory analysis of Time Series for hypothesizes feature values. In: International Scientific Conference UniTech 2017, vol. II, pp. 399-403, University publishing house V. Aprilov, Gabrovo (2017)

2 Balabanov T.D., Blagoev I.I., Dineva K.I. Self Rising Tri Layers MLP for Time Series Forecasting. In: Vishnevskiy V., Kozyrev D. (eds) Distributed Computer and Communication Networks. DCCN 2018. Communications in Computer and Information Science, vol 919. Springer, Cham. https://doi.org/10.1007/978-3-319-99447-5_50 (2018)

3 Balabanov, T., Atanasova, T., Blagoev, I., Activation Function Permutation for Multilayer Perceptron Training, International Conference on Big Data, Knowledge and Control Systems Engineering BdKCSE'2018, Sofia, Bulgaria, ISSN 2367-6450, pp. 9-14 (2018)

4 Blagoev I., Dokev N.: A Method for Investigating the Alterations in the Price Trends of the Currency Markets and Forecasting of Probable Future Alterations, *Problems of Engineering Cybernetics and Robotics*, vol.65, pp.39-48 (2012)

5 Blagoev I., Neglected Cybersecurity Risks in the Public Internet Hosting Service Providers. *Information&Security International Journal - ISIJ*, 47, no. 1, pp. 62-76 https://doi.org/10.11610/isij (2020)

6 Blagoev I.: Method for Evaluating the Vulnerability of Random Number Generators for Cryptographic Protection in Information Systems. In: Dimov I., Fidanova S. (eds) Advances in High Performance Computing. HPC 2019. Studies in Computational Intelligence, vol 902. Springer, Cham. https://doi.org/10.1007/978-3-030-55347-0_33. (2021)

7 Blagoev, I., Using R Programming Language for Processing of Large Data Sets, Proc. Int. Conf. Big Data, Knowledge and Control Systems Engineering – BdKCSE'2018, 21-22 November 2018, Sofia, Bulgaria ISSN 2367-6450, pp. 91-98.

8 Camara C., Martín H., Peris-Lopez P., Aldalaien M., Design and Analysis of a True Random Number Generator Based on GSR Signals for Body Sensor Networks, *Sensors* 19, 2033; doi:10.3390/s19092033 (2019)

9 Plummer T., Forecasting Financial Markets: The Psychology of Successful Investing, January (2010)

10 Pseudo-Random Number Generators, https://crypto.stanford.edu/pbc/notes/crypto/prng.html

11 Zhao P., R with Parallel Computing from User Perspectives, https://www.r-bloggers.com/r-with-parallel-computing-from-user-perspectives/ (2016)

12 Brown R. G.: Dieharder: A Random Number Test Suite, https://webhome.phy.duke.edu/~rgb/General/dieharder.php (2021)

13 Ciampi F., G. Marzi, S. Demi, M. Faraoni, The big data-business strategy interconnection: a grand challenge for knowledge management. A review and future perspectives, *Journal of Knowledge Management,* Vol. 24, Issue 5 (2020).
14 Koeune F. Pseudo-random number generator. In: van Tilborg H.C.A. (eds) Encyclopedia of Cryptography and Security. Springer, Boston, MA . https://doi.org/10.1007/0-387-23483-7_330 (2005)
15 Borcard, D., Gillet, F., Legendre, P. Numerical Ecology with R, Springer, pp. 9 – 30 (2011)
16 Baumer B. S., Kaplan D. T., Nicholas J., Modern Data Science with R, Horton Chapman & Hall/CRC, Boca Raton, (2017)
17 Long C. (Ед.) Data Science & Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, John Wiley & Sons, Inc., (2015)
18 Martínez-Acosta L., Medrano-Barboza J.-P., López-Ramos Á., López J., López-Lambraño Á., SARIMA Approach to Generating Synthetic Monthly Rainfall in the Sinú River Watershed in Colombia, *Atmosphere,* 11, 602; doi:10.3390/atmos11060602 (2020)
19 Mikalef P., Krogstie J., Examining the interplay between big data analytics and contextual factors in driving process innovation capabilities, *European Journal of Information Systems,* Volume 29, - Issue 3: Business Process Management and Digital Innovation https://doi.org/10.1080/0960085X.2020.1740618 (2020)
20 Scott G., Carr M., Cremonie M., Technical Analysis: Modern Perspectives, e CFA Institute Research Foundation (2016)
21 The R Journal, ISSN: 2073-4859, https://journal.r-project.org/ (2017)
22 Tomov, P., Monov, V., Artificial Neural Networks and Differential Evolution Used for Time Series Forecasting in Distributed Environment, Proc. of Int.conference Automatics and Informatics, ISSN 1313-1850, pp.129-132, Sofia, Bulgaria, (2016)
23 Wafi A.S., Hassan H., Mabrouk A., Fundamental Analysis Models in Financial Markets – Review Study, *Procedia Economics and Finance*, Vol. 30, 939 – 947. Elsevier (2015)
24 Wang, W., Y. Wang, Analytics in the era of big data: The digital transformations and value creation in industrial marketing, *Industrial Marketing Management,* Vol. 86, pp. 12-15, ISSN 0019-8501, https://doi.org/10.1016/j.indmarman.2020.01.005 (2020)
25 Li C., Zhang J., Sang L., Gong L., Wang L., Wang A., Wang Y., Deep Learning-Based Security Verification for a Random Number Generator Using White Chaos, *Entropy,* 22, 1134; doi:10.3390/e22101134 (2020)