УДК 003.26.09, 004.056.55, 519.25
Ключевые слова: Статистический крипто анализ, хеш-функция, SHA-256, Биткоин, частичная коллизия

Анализпрообразов функции SHA-256, хеши которых имеют особый вид

Аннотация

Многие современные электронные криптовалюты в своих протоколах используют хеш-функцию SHA-256. В этих протоках указано, что каждый блок транзакций подписывается хешем особого вида, все эти хеши совпадают в первых nбитах. В связи с увеличением популярности электронных валют, резко увеличился хешрейт сети, и начиная с 2017 года резко возросли и требования к хешу, подписывающему блок транзакций. Например, за 2018 год получены десятки тысяч прообразов, чьи хеши имеют одинаковую левую часть (70 из 256 битов). Фактически это означает, что непрерывно начиная с 2010 года проводится атака на функцию SHA-256, в плане подбора частичной коллизии. По количеству задействованных вычислительных ресурсов эта атака является самой масштабной за всю историю криптографии. В работе представлены результаты статистического анализа таких прообразов. Проверены гипотезы о нормальности распределения числа битов для каждой позиции в прообразе, числа встречаемости всех возможных комбинаций битов стоящих в любых двух, а также в любых трех позициях прообраза.

Введение

В настоящее время для обеспечения функционирования криптовалют за действованы колоссальные вычислительные ресурсы. Главной идеей проведенного исследования является то, что результаты работы этих сетей можно использовать для криптоанализа алгоритмов, на которых базируется протоколы этих электронных денежных систем.

Алгоритм SHA-256 был разработан Агентством национальной безопасности США(АНБ) и опубликован в августе 2002года[1]. Этот алгоритм описывает одностороннюю хеш функцию, аргументом которой является битовая строка произвольной длинны, а значением функции – строка длинной 256 битов. Функция односторонняя, так как легко вычисляется значение функции для любого аргумента, но получить прообраз, для заданного значения – сложно. В настоящий момент не существует известного алгоритма получения прообраза, который выполнит работу быстрее, чем с помощью полного перебора. Другим требованием, предъявляемым к хеш функциям, является отсутствие коллизий, то есть кроме как полным перебором нельзя найти два различных прообраза, имеющих один и тот же образ. В настоящий момент алгоритм SHA-256 используется повсеместно, например, при установлении защищенного соединения с сайтом.

Алгоритм SHA 256 состоит из 64-х итераций. В марте 2008 года индийские исследователи Сомитра Кумар Санадия и Палаш Саркар опубликовали найденные ими коллизии для 22 итерацийэтого алгоритма [2] и хотя SHA-256 формально не взломан, но в 2012 году появилось следующее поколение хеш-функций:SHA-3.

До этого момента, в 2008 году математиком или группой математиков под псевдонимом Сатоси Накамото был опубликован протокол и принципы работы новой электронной валюты – Биткоин. В нем все движения денежных средств этой валюты то есть транзакции - предлагалось упаковывать в блоки. В каждый отдельный момент времени активен только один (новый блок). Блок открывается, в него собирается информация о текущих перемещениях электронных денег между счетами. Приблизительно через 10 минут активности блока, он должен быть закрыт и подписан композицией хеш функцииSHA-256. После этого открывается новый блок.

Интересен способ, которым обеспечивается время активности блоков: как таковой жесткой привязки к 10-ти минутному промежутку нет, блок закрывается, когда для него найден особенный хеш. Каждый блок имеет

заголовок, в котором кроме прочей информации, содержится поле «Nonce». В этом поле хранится произвольное число и, соответственно, меняя значение «Nonce» можно получить множество различных хешей одного и того же блока. Так как общая идея любой криптовалюты заключается в децентрализации (каждый узел сети независимо от других работает для поддержания актуальной информации о состоянии электронных счетов), то каждый из узлов имеет возможность предложить хеш, который бы закрыл текущий блок транзакций. Но из всего множества генерируемыххешей нужен только один. Такой хеш называется **валидным** и определяется протоколом Биткоин как хеш, который меньше некоторого заранее заданного числа, называемого **целью**. В силу того, что хеш функцияSHA-256 односторонняя, то поиск такого хеша производится простым перебором числа «Nonce».Цель меняется в зависимости от количества генерируемых сетью хешей. Если узлов в сети становится много, блоки начинают закрываться быстрее чем за 10 минут, то цель уменьшается, сложность нахождения валидных хешей возрастает. И наоборот.

Этот принцип называется **доказательством работы**, потому что узел, нашедший валидный хеш, получает солидное вознаграждение (в электронной валюте).

Из валидности хеша следует, что такие хеши являются битовыми последовательностями, у которых несколько первых битов являются нулевыми. Если первые блоки биткоина имели валидные хеши, содержащие только 32 первых нулевых битов, то в настоящее время валидные хеши имеют72 первых нулевых битов. Наличие хешей, у которых совпадает какая то часть, можно назвать частичной коллизией.

Резкий рост курса биткойна в конце 2017 года, вызвал появление большого количества узлов, подбирающих хеши, и соответственно инициировал уменьшение значения цели. Например, в январе2019г. хешрейт сети биткоин составлял 42 эксахеша. Чтобы найти один валидный хеш, каждые 10 минут перебиралось порядка 25 зетта хешей, поэтому можно рассматривать работу биткойн-сети, как самую масштабную когда либо проводимую криптоатаку. Криптоатакой это является хотя бы потому, что полученные валидные хеши и их прообразы сохраняются и могут быть использованы для статистического анализа. Кроме того, при таком переборе хешей всегда остается вероятность нахождения коллизии.

Таким образом, сложилась ситуация двоякая ситуация: с одной стороны, алгоритм SHA-256 был разработан АНБ и его специалисты, конечно, провели тщательный анализ на устойчивость к криптоатакам. Алгоритм существует уже 17 лет и был протестирован не только известными группами криптографов, но и наверняка криптографами всех спецслужб мира. То, что алгоритм продолжает повсеместно использоваться, говорит о том, что он не был взломан.

С другой стороны, хотя SHA-256 используется в алгоритме многих криптовалют, он не предназначался для «доказательства работы» и изначально не тестировался на возможность получения частичных коллизий. Кроме того, аппаратных ресурсов, которые задействованы в сети биткоин сейчас, нет, и не было ни у одной страны мира.

Результаты работы сети Биткоин за последний год, открывают возможность для поиска алгоритма, с помощью которого можно было бы находить частичные коллизии, быстрее, чем с помощью полного перебора.Наличие или отсутствие такого алгоритма существенным образом отразится на принципах функционирования многих криптовалют.

В данной работе проверяется гипотеза: хеши, которые совпадают в значительной части своих первых битов, должны иметь прообразы, имеющие какой-то общий признак. Этот признак может быть и вероятностным, то есть прообразы, имеющие этот признак должны стохастически чаще иметь валидные хеши.

План исследования

Каждый блок биткойна содержит заголовок и хеш, полученный путем применения к нему функции SHA-256 дважды, т.е. SHA256(SHA256(Заголовок)). Для получения анализируемой выборки были обработанызаголовки блоков, сформированные с декабря 2017 по январь 2019, то есть в период, когда валидные хеши стали иметь 72 и более битовых нуля. Количество таких блоков **57615** и для каждого

заголовка был получен его хеш, т.е. применена функция SHA256(Заголовок). Эти хеши сами являются прообразами хешей SHA256(SHA256(Заголовок)), то есть прообразами валидных хешей. Так как каждый прообраз является последовательностью из 256 битов, то все полученные прообразы представляют собой битовую таблицу из 57615 строк и 256 столбцов.

Большой интерес представляет вопрос о том, расположены ли биты в этой таблице равновероятно. Если для каких-то битовых позиций в хеше будет установлен другой закон распределения, либо при анализе строк будет найдена зависимость значений битов каких то позицийот значений в других позицияхэтой же строки, то эти результаты можно будет использовать при нахождении валидных хешей быстрее, чем с помощью полного перебора. Заметим, что каждая строка соответствует хешу, полученному в какой-то момент времени, поэтомудля проводимого исследования порядок строк не важен. Кроме того, нет цели найти какую-то особенную строку или группу строк. Обратная ситуация с колонками – искомый общий признак должен объединять колонки с определенными номерами. Правило должно быть справедливо для всех строк либо таких строк должно быть достаточно много, чтобы можно было говорить о вероятностном признаке. Из этих соображений при анализе колонок не использовались статистические критерии, каждая колонка (или группа колонок) рассматривалась отдельно. При анализе строк, наоборот, данные по строкам были объединены в выборку, которая и была обработана с помощью статистического критерия.

Анализ количества единичных битов в столбцах

Рассмотрим произвольную колонку таблицы и изучим количество единичных битов в ней. Если биты распределены равновероятно, то количество единичных (а, соответственно, и нулевых) битов является суммой, в которой каждое слагаемое имеет распределение Бернулли с вероятностью успеха 0,5, следовательно,изучаемая сумма имеет биномиальное распределение с параметрами $n = 57615$ и $p = 0{,}5$. Так как слагаемых достаточно много — 57615, то в силу центральной предельной теоремы эту сумму можно считать нормально распределеннойс математическим ожиданием $M = np = 28807{,}5$и дисперсией $D = np(1-p) = 14403{,}75$.

Таким образом, нулевая гипотеза состоит в том, что количество единичных битов в выбранной колонке нормально распределено с указанными параметрами, то есть достаточно близко к величине $14403{,}75$.

Заметим, что если в рассматриваемой n-ой колонке единичные биты отсутствуют (либо наоборот, присутствуют в каждой строке), то можно говорить о наличии невероятностной зависимости. В этом случае, скорость поиска валидных хешей увеличится в 2 раза, так как прообразы у которых в n-ой позиции находится единичный бит (либо наоборот нулевой) – можно сразу отбрасывать, не беря от него хеш.

Если таких колонок обнаружится k-штук, то скорость поиска возрастет в 2^kраз.

В проведенном исследовании были проанализированы все 256 колонок, в каждом случае найдено отклонение от математического ожидания числа единичных битов от математического ожидания M.

Известно, что нормально распределенная случайная величина с вероятностью 0,9973не отклоняется от своего среднего значения на3σ (правило 3х сигм). В нашем случае $\sigma = \sqrt{np(1-p)} = 120{,}02$.

Если существует невероятностная зависимость, то в таких колонках отклонение достигнет

$$\frac{57615 - M}{\sigma} = 240\sigma$$

В таблице 1 приведены отклонения наблюдаемых значений от математического ожидания M.

Таблица 1. Отклонения наблюдаемых значений от математического ожидания M, выраженные в величинах σ

0,77	1,07	0,91	0,61	1,37	1,21	2,11	0,34	0,71	0,02	1,08	0,82	1,47	1,10	1,98	2,02
1,34	0,8	0,24	1,25	0,97	0,77	0,51	0,29	0,02	0,36	0,92	0,50	0,69	1,29	1,44	0,35
0,64	1,38	0,62	0,25	1,65	1,02	1,52	0,02	0,03	2,21	0,99	0,33	0,38	1,47	1,37	1,54
0,17	0,27	1,22	0,55	0,36	0,94	1,71	0,02	0,11	0,21	0,46	0,71	0,42	1,46	2,59	0,92
0,44	1,55	0,55	0,55	0,57	1,37	0,72	1,46	0,42	0,86	0,04	0,18	0,79	0,18	0,84	1,23
0,29	0,9	1,78	0,39	1,82	0,12	1,34	1,12	0,94	1,05	1,62	0,57	0,63	0,88	0,14	1,35
0,91	0,14	0,37	1,42	0,50	0,31	0,58	1,02	1,03	1,59	0,26	1,57	0,55	0,77	0,87	1,56
0,37	1,55	1,00	1,57	1,63	0,08	0,87	1,19	0,22	0,77	1,13	1,52	0,75	0,68	1,52	0,16
1,03	0,74	0,62	0,08	0,67	0,03	0,97	1,61	0,21	0,42	1,17	1,27	0,20	0,67	0,83	0,45
0,97	0,36	0,23	1,62	1,07	0,07	1,07	1,60	0,83	0,41	0,08	0,16	0,09	0,27	0,50	0,47
1,67	1,17	1,21	0,13	0,45	1,11	0,51	0,88	0,05	0,27	0,65	0,27	0,78	0,98	0,24	1,12
1,02	0,82	0,37	0,20	0,07	1,24	0,58	1,16	0,98	0,85	1,38	1,37	0,05	0,22	0,41	0,10
1,17	0,57	1,15	2,29	0,64	0,74	1,27	0,42	0,11	0,01	1,15	0,15	1,66	1,12	0,16	0,60
2,4	0,55	0,72	0,19	0,87	0,03	0,96	1,17	0,42	1,49	0,87	0,73	0,01	0,47	0,52	1,47
0,04	0,86	0,21	1,22	0,47	2,22	1,20	0,01	1,94	1,57	1,96	0,92	1,86	1,37	1,10	0,32
0,42	0,66	1,91	1,28	2,05	0,07	0,50	0,65	0,92	1,57	0,86	1,04	0,30	0,32	1,01	0,80

Первая строка содержит данные для колонок 1-16, вторая строка – для колонок 17-32 и т.д.

Наибольшее отклонение наблюдается в 63-й колонке, оно равно $2{,}59\sigma$, и если принимается нулевая гипотеза, то вероятность такого отклонения равна 0,009. Вероятность такого события достаточно мала, но следует учитывать то, что всего анализируемых значений 256, поэтому это отклонение могло быть действительно случайным.

Анализ пар столбцов

Рассмотрим произвольную пару столбцов анализируемой таблицы. Возможны следующие варианты значений в них: «11», «10», «01», «00». Если биты в этих столбцах распределены равновероятно, то вероятность встретить любой из этих элементов в наудачу взятой строке таблицы, равна p=1/4. Поэтому количество строк, содержащих любуюзаранее заданную битовую пару(«11», «10», «01», «00») в рассматриваемой паре столбцов, распределено по нормальному закону с математическим ожиданием $M = np$ =14403,75и дисперсией$D = np(1-p)$ =10802,81.

Если в какой-то колонке существует функциональная зависимость появления битовой пары, то отклонение от математического ожидания для такой колонки будет равно

$$\frac{57615 - M}{\sigma} = 416\sigma$$

При наличии k-штук таких пар колонок скорость поиска валидного хеша увеличится в 4^kраз.

В проведенном исследовании, были рассмотрены все возможные пары столбцов таблицы. Для каждой из них было подсчитано число содержащихся в них различных битовых пар («11», «10», «01», «00») и найдено отклонение этих чисел от математического ожидания. Нетрудно подсчитать количество полученных значений.Всего различных пар столбцов 32 640, для каждого столбца досчитывались 4 битовых пары, соответственно, всего оцениваемых значений: $4 \cdot 32640 = 130\,560$.

В 257 случаях наблюдалось отклонение величины от математического ожидания, более чем на 3 сигма, а в 33-х случаях – более чем на 3,5 сигма. Информация о таких колонках приведена в Таблице2.

Таблица 2. Отклонения наблюдаемых значений от математического ожидания M, более чем 3,5 σ

Отклонение (сигм)	Вероятность	NN колонок	Битовые пары
3,95	0,000079	230, 245	1, 0
3,90	0,000095	63, 125	1, 0
3,87	0,000107	62, 63	1, 1
3,86	0,000115	196, 230	0, 0
3,85	0,000118	63, 233	0, 0
3,82	0,000132	7, 243	0, 1
3,81	0,000140	7, 129	1, 0
3,80	0,000143	100, 243	0, 1
3,78	0,000155	9, 42	0, 0
3,66	0,000249	15, 42	0, 1
3,66	0,000249	17, 42	0, 1
3,64	0,000273	63, 173	0, 1
3,64	0,000273	63, 184	0, 0
3,63	0,000279	196, 209	0, 1
3,62	0,000300	17, 199	0, 0
3,62	0,000300	39, 129	1, 0
3,62	0,000295	116, 195	0, 1

Отклонение (сигм)	Вероятность	NN колонок	Битовые пары
3,60	0,000317	42, 196	0, 1
3,60	0,000317	91, 108	0, 0
3,60	0,000317	177, 245	0, 0
3,60	0,000317	209, 217	0, 1
3,59	0,000335	188, 245	0, 1
3,57	0,000361	136, 161	0, 1
3,56	0,000368	63, 235	0, 0
3,54	0,000395	218, 237	1, 0
3,53	0,000418	13, 245	1, 1
3,53	0,000410	15, 173	1, 1
3,53	0,000418	16, 243	0, 0
3,53	0,000418	91, 224	1, 1
3,53	0,000418	161, 243	1, 0
3,51	0,000449	16, 237	0, 1
3,51	0,000449	70, 248	1, 1
3,51	0,000441	127, 196	0, 1

Наибольшее отклонение наблюдается для битовой пары «1,0», оно составило $3{,}95\sigma$. Если считать, что биты в таблице распределены равновероятно, то вероятность такого отклонения равна 0,000079. Учитывая общее количество анализируемых значений, скорее всего эти отклонения случайны.

Анализ троек столбцов

Рассмотрим произвольную тройку столбцов анализируемой таблицы. Возможны следующие варианты значений в них: «111», «110», «101», «100», «011», «010», «001», «000». При равновозможном размещении битов вероятность встретить в произвольно выбранной строке любую из этих битовых троек равна 1/8. Аналогично предыдущим случаям, число строк, содержащих заранее заданную битовую тройку, подчинено нормальному закону распределения.Математическое ожидание этого распределения равно $M = np = 7201{,}875$, дисперсия - $D = np(1-p) = 6301{,}641$.

Наибольшее возможное отклонение математического ожидания от наблюдаемого значения (в случае неслучайности расположения битовых троек) равно

$$\frac{57615 - M}{\sigma} = 635\sigma$$

При наличии k-штук таких троек колонок скорость поиска валидного хеша увеличится в 8^k раз.

В проведенном исследовании, были рассмотрены все возможные тройки столбцов таблицы. Для каждой из них было подсчитано число содержащихся в них различных битовых пар и найдено отклонение этих чисел от математического ожидания. Всего было обработано $2763520 \cdot 8 = 22\ 108\ 160$ значений. Отклонения от 3-х сигма наблюдалось в 54 376 случаях, из них отклонения более 4 сигма – в 1049 тройках, а более 5 сигма – только в трех:

Таблица 3. Отклонения наблюдаемых значений от математического ожидания M, более чем 5 σ

Отклонение (сигм)	Вероятность	NN колонок	Битовые тройки
5,69	10^{-8}	112, 209, 217	0, 0, 1
5,01	$5{,}32 \cdot 10^{-7}$	7, 80, 243	0, 0, 1
5,00	$5{,}68 \cdot 10^{-7}$	196, 209, 230	0, 1, 0

Вероятность таких отклонений достаточно мала, но и выборка велика. Можно оценить вероятность появления таких отклонений в данной выборке: при проверке любой тройки столбцов, превышение математического ожидания более или равным 5,69σ будем считать успехом, вероятность успеха равна $p_y = 10^{-8}$. Испытания независимые, соответственно, подчинены закону Бернулли. Вероятность успеха мала, число испытаний $n_и = 22\,108\,160$ – велико, соответственно, вероятность появления одного успеха может быть аппроксимирована распределением Пуассона с параметром $\lambda = n_и p_y = 0{,}22$ и значением$P(\lambda) = 0{,}177$. Как видим, это событие достаточно вероятно, поэтому и в этом случае отвергать гипотезу о равновероятном распределении битов нельзя, но возможно в этом случае необходимо провести дополнительные исследования.

Изучение комбинаций из четырех столбцов затруднено большой вычислительной сложностью. Всего возможных четверок столбцов – 349 585 280, а возможных битовых четверок – 16, соответственно, общее количество исследуемых величин равно 5 593 364 480, что в 253 раза больше, чем при анализе троек столбцов. Предыдущий анализ при обработке на персональном компьютере, потребовал порядка 5 часов машинного времени, анализ четверок столбцов потребует уже около 52 суток.

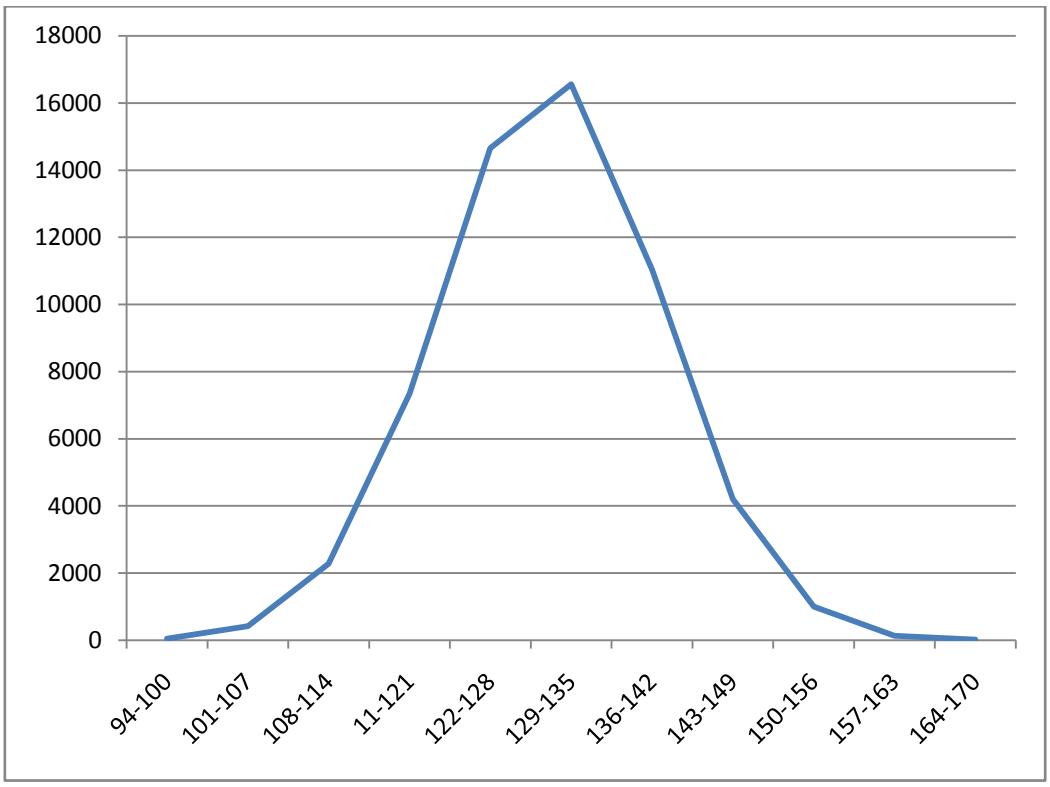
Анализ количества единичных битов в строках (горизонтальный анализ)

Для каждой строки подсчитаем количество единичных битов. Если биты распределены равновероятно, то количество единичных (а, соответственно, и нулевых) битов является суммой, в которой каждое слагаемое имеет распределение Бернулли с вероятностью успеха 0,5, следовательно,изучаемая сумма имеет биномиальное распределение с параметрами $n = 256$ и $p = 0{,}5$. Эту сумму можно считать нормально распределенной случайной величиной с математическим ожиданием $M = np = 128$ и дисперсией $D = np(1-p) = 64$. Так как каждая строка это некоторый прообраз, то для нас не важен порядок их следования (мы не ищем различий между хешами, а наоборот – некий общий признак), а значит, все полученные суммы можно объединить в выборку и исследовать ее на нормальность.

Нулевая гипотеза состоит в том, что полученная выборка имеет нормальное распределение.

Для проверки этой гипотезы использовался критерий Пирсона χ^2. Всего элементов выборки (количество строк таблицы) – 57615, Наибольшее наблюдаемое значение: 163, наименьшее – 97, выборка была разбита на 11 интервалов длиной 6.

Рис 1. Графическое изображение вариационного ряда



При проверке гипотезы использовался уровень значимости: α=0,05; критическое значение критерия: $\chi^2_{\text{крит}} = 15{,}5$. Полученное наблюдаемое $\chi^2_{\text{набл}} = 6{,}38$,следовательно, нет оснований отвергать гипотезу о нормальном распределении числа единичных битов в строках таблицы, а значит гипотезу о равновероятном распределении битов в таблице отвергать нельзя.

Выводы и дальнейшие перспективы исследования

На основе данных, полученных из блоков криптовалюты Биткоин, была составлена таблица прообразов валидных хешей. Для поиска некоторого признака, обобщающего эти прообразы был проведен ряд исследований. Вертикальный анализ состоял в изучении количества единичных битов в каждой колонке, в каждой паре и тройке колонок. Горизонтальный анализ заключался в проверке соответствия числа единичных битов в строках нормальному распределению.

Хотя не было выявлено явных признаков того, что единичные биты распределены в таблице не равновероятно, но результаты, полученные при анализе троек колонок интересно было бы проверить на других данных, например, полученных из блоков Биткоин, сгенерированных в течение 2019 года.

Заметим, что данная работа не охватывает всего множества возможных тестов полученной таблицы. Интересно было бы расширить горизонтальный анализ, изучив не только количество единичных битов в строках, но и количество различных цепочек битов. Дальнейший анализ колонок предполагает изучение всех четверок колонок (возможно на распределенной вычислительной системе).

Список использованных источников

1. WouterPenard, TimvanWerkhoven. On the Secure Hash Algorithm family // https://web.archive.org/web/20160330153520/http://www.staff.science.uu.nl/~werkh108/docs/study/Y5_07_08/infocry/project/Cryp08.pdf (датаобращения: 29.03.2019)

2. Somitra Kumar Sanadhya, PalashSarkar. 22-Step Collisions for SHA-2 // https://eprint.iacr.org/2008/270.pdf (датаобращения: 29.03.2019)