УДК 004.056.55

Сагун Андрій Вікторович
кандидат технічних наук, доцент, доцент кафедри комп'ютерних систем, мереж та кібербезпеки,
Національний університет біоресурсів і природокористування України
ORCID: https://orcid.org/0000-0002-5151-9203
E-mail: a.sagun@nubip.edu.ua

Місюра Максим Дмитрович
кандидат технічних наук, доцент, доцент кафедри комп'ютерних систем, мереж та кібербезпеки,
Національний університет біоресурсів і природокористування України
ORCID: https://orcid.org/0000-0002-9061-3462
E-mail: mdm@nubip.edu.ua

Білич Мілана Сергіївна
бакалавр, Національний університет біоресурсів і природокористування України
E-mail: kib22-m.bilych@nubip.edu.ua

ТЕХНОЛОГІЯ АНАЛІЗУ ЯКОСТІ ХАРАКТЕРИСТИК ГЕШ-ФУНКЦІЇ

Анотація. *Запропонована технологія аналізу якості геш-функції. Для перевірки якої спеціально розроблена геш-функція, яка є спрощеним аналогом MD5. Показано, що традиційна оцінка якості отриманих геш-значень на основі наявності кількості колізій має суттєвий недолік - висока обчислювальна складність. Тому, технологія оцінки якості зводиться до апроксимації аналітичного виразу функції, який піддається математичному аналізу традиційними методами, застосовними до поліномів. Більш якісною геш-функцією вважатися та геш-функції, гістограма якої є максимально наближеною до графіку функції y=x. Це буде відповідати гаусівському розподілу величин. При використанні запропонованої технології необхідно розробити критерії класифікації функцій за відхиленнями від ідеального розподілу значень за одним або декількома з таких показників, як: дисперсія; математичне очікування; середнє арифметичне; середнє геометричне значення тощо.*

Ключові слова. *Геш-функція, колізії, інтерполяція функції, апроксимація функції, якість геш-функції*

Вступ. Відомо, що основним параметром, за яким характеризується та чи інша геш-функція є кількість колізії, наявність чи відсутність колізії [1,2]. Виявлення колізії для ідеального випадку дослідження якості геш-функції має проводитися для великого діапазону вихідних значень $H(x)$. Тільки в такому випадку результати дослідження можуть визнаватися статистично значущими [1-3]. Ймовірність колізії в геш-функції часто оцінюється за парадоксом днів народження, який базується на кількості можливих вихідних значень N (розмірність хешу) [4,5]. Даний підхід є найбільш оптимальний з точки зору витраченого часу і прийнятний стосовно точності отриманого показника якості функції.

Тому, підхід оцінки якості функцій за методом навіть часткового перебору, навіть, з огляду на його визнану високу точність і прийнятність отриманих результатів, часто складно застосувати через занадто високі часові витрати на реалізацію.

Аналіз графічних відображень результатів отриманих вихідних значень геш-функції є важливим для оцінки рівномірності та випадковості її поведінки. Хоча, загалом, частіше використовують математичні методи дослідження геш-функцій, але включення до методик оцінювання якості геш-функцій результатів візуалізації вихідних значень значно доповнює і унаочнює статистичний аналіз отриманого статистичного розподілу.

Мета дослідження – створення та вдосконалення методів візуалізації та аналізу графічного представлення вихідних даних геш-функцій, що дозволить оцінювати їх криптографічні властивості, такі як аваланч-ефект, рівномірність розподілу та стійкість до колізій.

Результати дослідження. Припустимо, що ми маємо геш-функцію виду $H(x)$, яка повертає вихідне значення фіксованої довжини. Для математичного аналізу можна використати наступні результати графічної візуалізації її значень [6]:

1. Графік залежності значення хешу від вхідних даних. Для певного набору вхідних значень x (наприклад, $x = 1,2,3$), обчислюємо відповідні значення $H(x)$.

Отримані значення функцій можливо візуалізізувати у вигляді точкового графіку або гістограми.

Далі з'являється можливість перевірити рівномірність отриманого розподілу значень функції $H(x)$.

2. Аналіз біткарт (Bit Distribution Analysis). Такий вид графічного аналізу передбачає, що кожен вихідний хеш можна порівняно легко представити у вигляді двовимірного бітового зображення (наприклад, 1024-бітні значення у вигляді матриці розмірністю 32×32).

В такому випадку графічна візуалізація отриманих результатів гешування дає змогу перевірити, чи існують ти чи інші закономірності у вихідних бітових послідовностях.

3. Графік змін у вихідному значенні при малих змінах вхідних даних (Avalanche Effect Analysis). При такому виді дослідження графіку вихідних значень геш-функції для кожного вхідного значення x обчислюємо значення функції $H(x)$ і $H(x+1)$, а потім вимірюємо отриману різницю в бітах.

Висновок стосовно якості отриманої робиться на основі характеру зміни вхідних даних, в залежності від зміни вхідних значень:

- геш-функція вважається якісною, якщо зміна хоча б одного біта у вхідних даних призводить до змін у значній кількості бітів вихідного значення (ефект лавини).
- в іншому випадку геш-функція $H(x)$ вважається неякісною.

Для ілюстрації роботи даного методу будується графік, що показує кількість змінених бітів у хеші для кожної зміни у вхідному повідомленні.

Для прикладу розробимо геш-функцію з такими характеристиками:

1) розрядність вихідних значень функції – 128 біт;

2) на вхід такої функції може подаватися будь-яка послідовність, довжиною до 1 Мбіт (1 048 576 біт);

3) підготовка вхідної послідовності з доповнення реалізуємо аналогічно до описаного у відкритих джерел для геш-функції md5 [7].

5) механізм «перемішування» (permutation) даних в розроблюваній геш-функції реалізуємо по аналогії з алгоритмом функції md5 [7].

6) обрахуємо параметри якості отриманої геш-функція по таким параметрам: дисперсія, математичне очікування тощо на широкій виборці вихідних значень типу «геш-128» та побудуємо відповідні ілюстративні графіки якісних параметрів хеш-функції.

Отримаємо результати оцінки якості геш-функції з використанням методів математичної статистики:

- математичне очікування: 9410810830297161728
- дисперсія: 28080520602751215165548159324164980736
- кількість колізій: 0 (на вибірці з 850 тестів).

На основі отриманих даних побудуємо гістограму розподілу значень створеної геш-функції. На гістограмі по осі X будуть відображатися отриманні геш-значення (64-бітові числа), а по осі Y – значення статистичного показника частоти появи певних геш-значень серед випадкових тестових вхідних 1024 бітових повідомлень, які генеруються програмною реалізацією алгоритму розробленої геш-функції (рис.1).

Наведена на рис. 1 гістограма дозволяє візуально оцінити дисперсію отриманих 850 значень геш-функції. Хоча для такої кількості вхідних значень колізій знайдено не було, але повний цикл пошуку колізій мав би охопити доволі багато значень. Так, для знаходження колізії потрібно перебрати всі можливі вихідні значення. Оскільки вихід має 128 біт, то кількість можливих унікальних гешів складе: $2^{128} \approx 3,4 \times 10^{38}$. Це означає, що в гіршому випадку потрібно перевірити $3,4 \times 10^{38}$ унікальних вхідних значень, щоб гарантовано знайти

колізію. При використанні атака типу «день народження» [8] можна зменшити перебір до $2^{64} \approx 1{,}8 \times 10^{19}$ ітерацій. Це значно швидше, але все одно дуже обчислювально складно.

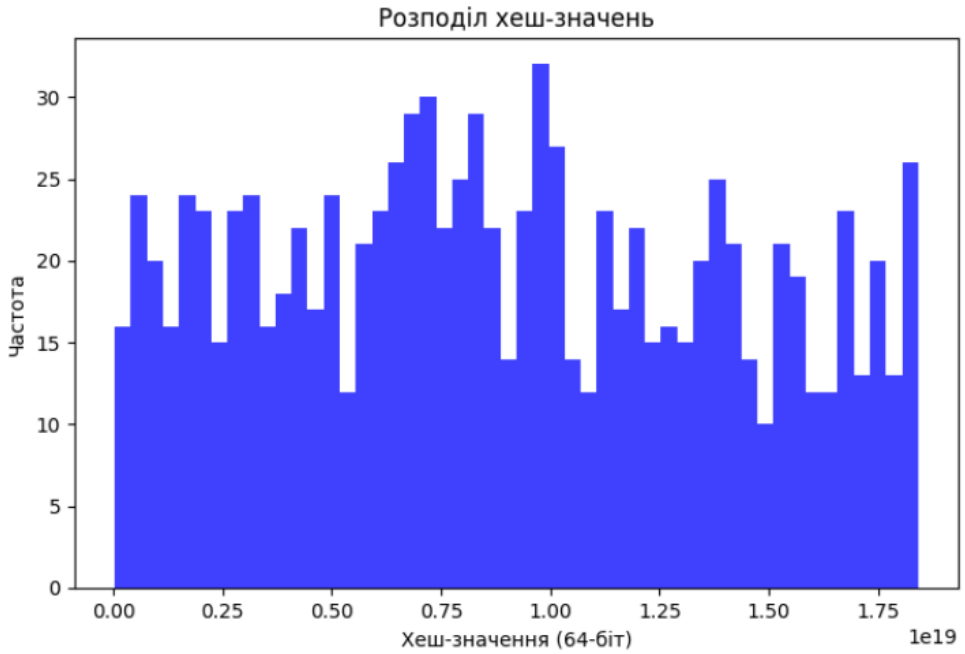

Рисунок 1 – Гістограма частоти розподілу вихідних значень створеної геш-функції в першій ітерації

Якщо для перебору всіх можливих значень аргументу розробленої функції використати сучасну відеокарту, наприклад NVIDIA RTX 4090. То, для відомого і співставного з розробленим за складність алгоритму MD5 (має 128-бітний вихід) обчислювальна потужність дозволяє обчислювати 200-300 млрд гешів на секунду (тобто 2×10^{11} до 3×10^{11}хешів/сек) для перебору всіх можливих значень потрібно від 1.13×10^{27}секунди $\approx 3{,}6 \times 10^{19}$років.

За умови здійснення більш оптимальної атаки «дня народження» обчислювальна складність становитиме: $2^{64} \approx 1.8 \times 10^{19}$ ітерацій, а час обчислення складе: 6×10^{7} секунд =1,9 років.

Відновлення аналітичного виразу функції та аналіз її графічної характеристики.

Математичний метод відновлення функції за відомими координатами точок називається інтерполяція або апроксимація. Отже, маючи графік вихідних значень геш-функції можна отримати аналітичний вираз даної функції та піддати її математичному аналізу.

Маємо потребу відновити функцію точно, то можна використовувати інтерполяцію. В іншому випадку використовують апроксимацію (якщо відомо, що існуючі дані містять похибки). Через те, що вихідні результати похибок не мають, то використовуємо апроксимацію.

Для отримання аналітичного виразу даної функції використаємо координати точок, отриманих за результатами попередніх розрахунків, що відповідають графіку, наведеному на рис. 1.

Побудуємо аналітичний вираз функції, використовуючи інтерполяційний поліном Лагранж, сплайн-інтерполяцію або метод найменших квадратів [9]. За результатами проведеним спроб, метод Лагранжа виявився нестабільним через велику кількість точок. Тому, використовуємо інтерполяцію цих точок за допомогою поліноміальної апроксимації. Для цього використаємо метод найменших квадратів для знаходження полінома, який найкраще підходить до цих даних [9].

Апроксимована функція для отриманих значень гістограми для 1 випадку:

$$y_1 = 0.0002x^3 - 0.0492x^2 + 3.0819x - 14.1699.$$

Апроксимована функція для отриманих значень гістограми для 2 випадку:

$$y_2 = -4.000866 \times 10^{-58}x^3 + 4.213750 \times 10^{-38}x^2 - 6.461849 \times 10^{-19}x + 21.80919.$$

Апроксимована функція для отриманих значень гістограми для 3 випадку:

$$y_3 = -3.89 \times 10^{-57}x^3 + 1.09 \times 10^{-37}x^2 - 7.02 \times 10^{-19}x + 20.19.$$

Після 5 повторів тестів генерування значення параметрів геш-функції на інтервалі були систематизовані і для кожного з набору значень було реалізовані спроби відновлення аналітичного виразу функції з використанням поліноміальної апроксимації, а не інтерполяції.

Вибір апроксимації продиктований тим, що необхідно відновити вираз, який "підходить" під більшість даних, тобто мінімізує загальну похибку по всіх відомих точках.

Це дозволяє отримати функцію, яка не обов'язково проходить через усі точки, але найкраще описує тенденції.

В той час, як інтерполяція є процесом знаходження полінома, який точно проходить через всі задані точки, а це, як видно з рис. 1, 2 і 3 є складною задачею.

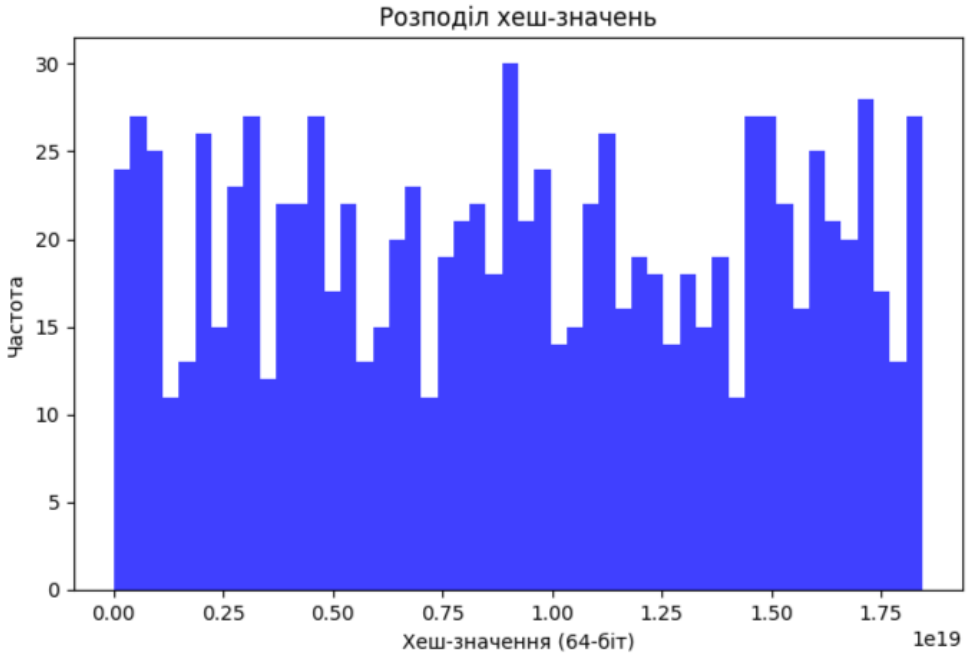

Рисунок 2 – Гістограма частоти розподілу вихідних значень створеної геш-функції в другій ітерації

Фактично, всі з наведених вище аналітичних функцій характеризують поведінку вихідного розподілу 128-бітних геш-значень розробленої функції $H(x)$.

Отже, математичний аналіз для випадку графічної характеристики геш-функції більш практичний. Це пов’язано з наявністю відповідних засобів математичного аналізу та значно меншим часом отримання результату перевірки на якість.

Формула для очікуваної кількості колізій при k випадкових хешах виглядає так:

$$P \approx 1 - e^{-\frac{k^2}{2N}}.$$

Якщо потрібно знайти таке значення параметру k, при якому кількість колізій стане більш помітною, то можна наступним виразом:

$$k \approx 1.2\sqrt{N},$$

де k — кількість гешів, які потрібно згенерувати для отримання хоча б однієї колізії з достатньо великою ймовірністю; $N = 2^N$ — загальна кількість всіх можливих значень хешу n-бітного хешу.

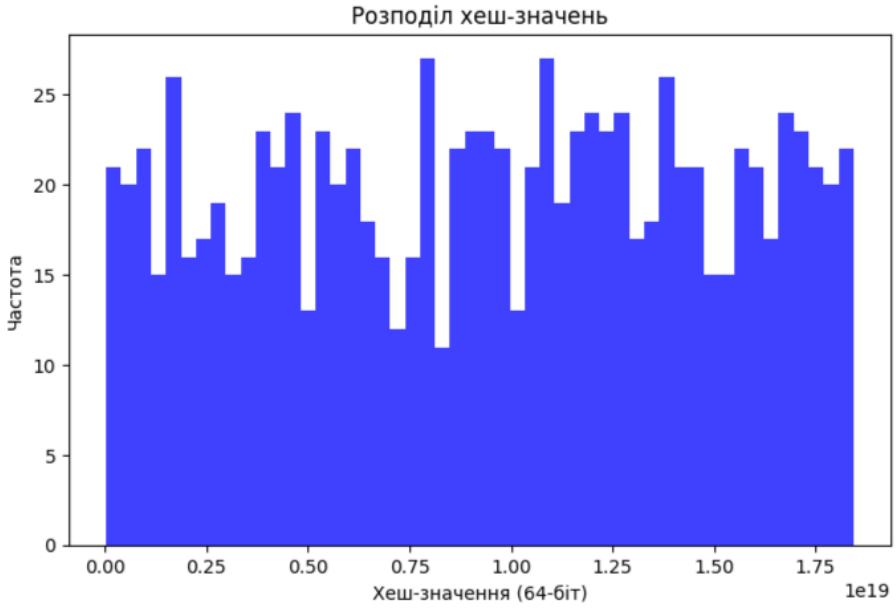

Рисунок 3 Гістограма частоти розподілу вихідних значень створеної геш-функції в третій ітерації

Для порівняння досліджувалася геш-функція MD5. Вона генерує на виході 128-бітні хеші, тобто $N = 2^{128}$.

Для такої функції параметр $k \approx 1{,}2\sqrt{2^{128}} = 1{,}2 \times 2^{64} \approx 2^{64}$. Щоозначає, що знадобиться приблизно 2^{64} різних хешів, для того, щоб отримати з високою ймовірністю хоча б одну колізію.

В той же час, відомо, що якщо потрібно отримати статистично значущу оцінку частоти колізій, то бажано для розробленої функції перевірити значно більшу вибірку [4-5], наприклад, 10^{20} або, навіть, більше хешів.

Після проведення аналізу програмної реалізації алгоритму геш-функції md5 (опис алгоритму взятий з відкритого джерела [10]) виявлялися 1000 повторень серед можливих 10^9 хешів. В результаті отримано емпіричний показник частоти колізій, який можна порівняти з теоретичними очікуваннями.

Цей показник приблизно співпадає з тим, що отримано для розробленої геш-функції (в межах статистичної помилки $\mu \in \pm(0 \dots 3)\%$.

Висновки та перспективи. В контексті оцінки якості практичну користь можуть давати не результати математичного аналізу відновленого за отриманими значеннями функції її аналітичного вигляду, а традиційні показники якості геш-функції, отримані з використанням математичної статистики, наприклад, такі, як:

- дисперсія;
- математичне очікування;
- середнє арифметичне;
- середнє геометричне значення тощо.

Однак, в такому випадку вимагає розробки критерії класифікації функцій за відхиленнями від ідеального розподілу значень за одним або декількома з таких показників.

Як показано в основні частині оцінки якості геш-функції но основі даних про колізії є занадто обчислювально складним і недоцільний для практичного застосування. З огляду на те, що для оцінки кількості колізій навіть відносно простої функції md5, потрібно проаналізувати принаймні $1{,}2\sqrt{N}$ значень хеш-функції, але для отримання більш точних даних необхідно збільшити цю кількість у декілька разів. Для всіх 128-бітних геш-функцій для цього потрібно провести приблизно $2^{64} \approx 18$ квінтильйонів значень. Тільки тоді можна отримати гарантовано високу ймовірність знаходження колізії. В той же час розглянутий альтернативний метод визначення якості геш-функції дозволяє зробити це на порядок швидше (залежно від структури алгоритму дослідженої функції вимагає додаткових досліджень).

З проведених вище досліджень можна побачити, що традиційний математичний аналіз інтерпольованих значень геш-функції для випадку графічної характеристики такої геш-функції є заскладненим і малозручним для практичного застосування. Хоча такий аналіз і може бути перспективний для деяких випадків (вимагає подальших досліджень).

Перспективи вдосконалення. Більш раціональним може бути візуальний швидкий аналіз отриманої гістограми якості. В такому випадку більш якісною буде вважатися та геш-функції, гістограма якої буде максимально наближено до графіку функції $y = x$. Це буде відповідати рівномірному розподілу значень (гаусівський розподіл) [11].

Для того, щоб створити умови для практичного застосування запропонованого методу оцінки слід провести додаткові дослідження, створити класи якості, ранжовані з допустимими від ідеальної функції відхиленнями.

Список використаних джерел

1. Wang, X., & Yu, H. (2005). Collisions of SHA-1: Second-preimage and preimage attacks. Crypto 2005 Rump Session. https://www.iacr.org/archive/crypto2005/36210017/36210017.pdf.
2. Rivest, R. (1992). The MD5 Message-Digest Algorithm (RFC 1321). Internet Engineering Task Force (IETF). https://doi.org/10.17487/RFC1321.
3. Tanasiuk, Yu. V., Melnychuk, Kh. V., & Ostapov, S. E. (2017). Rozrobka i doslidzhennia kryptohrafichnykh khesh-funktsii na osnovi klitynnykh avtomativ [Development and research of cryptographic hash functions based on cellular automata]. Systemy Obrobky Informatsii [Information Processing Systems], 4(150), 122–127.
4. Brassard, G., Høyer, P., & Tapp, A. (1998). Quantum cryptanalysis of hash and claw-free functions. In C. L. Lucchesi & A. V. Moura (Eds.), LATIN'98: Theoretical informatics (pp. 163–169). Springer. https://doi.org/10.1007/BFb0054319.
5. Bellare, M., & Rogaway, P. (2005). The birthday problem. In Introduction to modern cryptography (pp. 273–274). https://web.cs.ucdavis.edu/~rogaway/classes/227/spring05/book/main.pdf.
6. Horra, E., Beyene, A., & Yitagesu, S. (2024). Enhanced avalanche effect analysis algorithm considering both single and double key pair RSA algorithms (Preprint). Research Square. https://doi.org/10.21203/rs.3.rs-4113962/v1.
7. Telegin, V. (2023). Analysis of cryptographic strength of the modified MD5 algorithm. Universum: Tekhnicheskie Nauki [Universum: Technical Sciences], 114(9).
8. Goldberg, S. (1976). A direct attack on a birthday problem. Mathematics Magazine, 49(3), 130–132. https://doi.org/10.2307/2690270.
9. Shelevytskyi, I. V., Shutko, M. O., Shutko, V. M., & Kolganova, O. O. (2007). Splainy v tsyfrovii obrobtsi danykh i syhnaliv [Splines in digital data and signal processing].
10. Rivest, R. (1992). The MD5 Message-Digest Algorithm (RFC 1321). Internet Engineering Task Force (IETF). https://doi.org/10.17487/RFC1321.
11. Harasymchuk, O. I., & Maksymovych, V. M. (2003). Heheratory psevdownykhynykh chysel, yikh zastosuvannia, klasyfikatsiia, osnovni metody pobudovy i otsinky yakosti [Generators of pseudorandom numbers, their application, classification, main methods of construction and quality assessment]. Naukovo-tekhnichnyi zhurnal "Zakhyst informatsii" [Scientific and Technical Journal "Information Protection"], (3), 29–36.

Sahun Andrii
PhD, Associate Professor of Computer Systems, Networks and Cybersecurity Department
National University of Life and Environmental Sciences of Ukraine
ORCID: https://orcid.org/0000-0002-5151-9203
E-mail: a.sagun@nubip.edu.ua

Misiura Maksym
PhD, Associate Professor of Computer Systems, Networks and Cybersecurity Department
National University of Life and Environmental Sciences of Ukraine
ORCID: https://orcid.org/0000-0002-9061-3462
E-mail: mdm@nubip.edu.ua

Bilych Milana
bachelor, National University of Life and Environmental Sciences of Ukraine
E-mail: kib22-m.bilych@nubip.edu.ua

TECHNOLOGY FOR ANALYZING THE QUALITY OF HASH FUNCTION CHARACTERISTICS

Abstract. *The technology for analyzing the quality of a hash function is proposed. To test it, a specially developed hash function, which is a simplified analog of MD5, is used. It is shown that the traditional quality assessment of the obtained hash values based on the number of collisions has a significant drawback - high computational complexity. Therefore, the technology of quality assessment is reduced to the approximation of the analytical expression of the function, which is amenable to mathematical analysis by traditional methods applicable to polynomials. A better quality hash function is a hash function whose histogram is as close as possible to the graph of the function y=x. This will correspond to the Gaussian distribution of values. When using the proposed technology, it is necessary to develop criteria for classifying functions according to deviations from the ideal distribution of values by one or more of the following indicators: variance; mathematical expectation; arithmetic mean; geometric mean, etc.*

Keywords: *hash function, collisions, function interpolation, function approximation, quality of a hash function*