Evaluation functions in the cryptoanalysis of homophonic substitution

Eugen Antal, Marek Sys, Juraj Varga

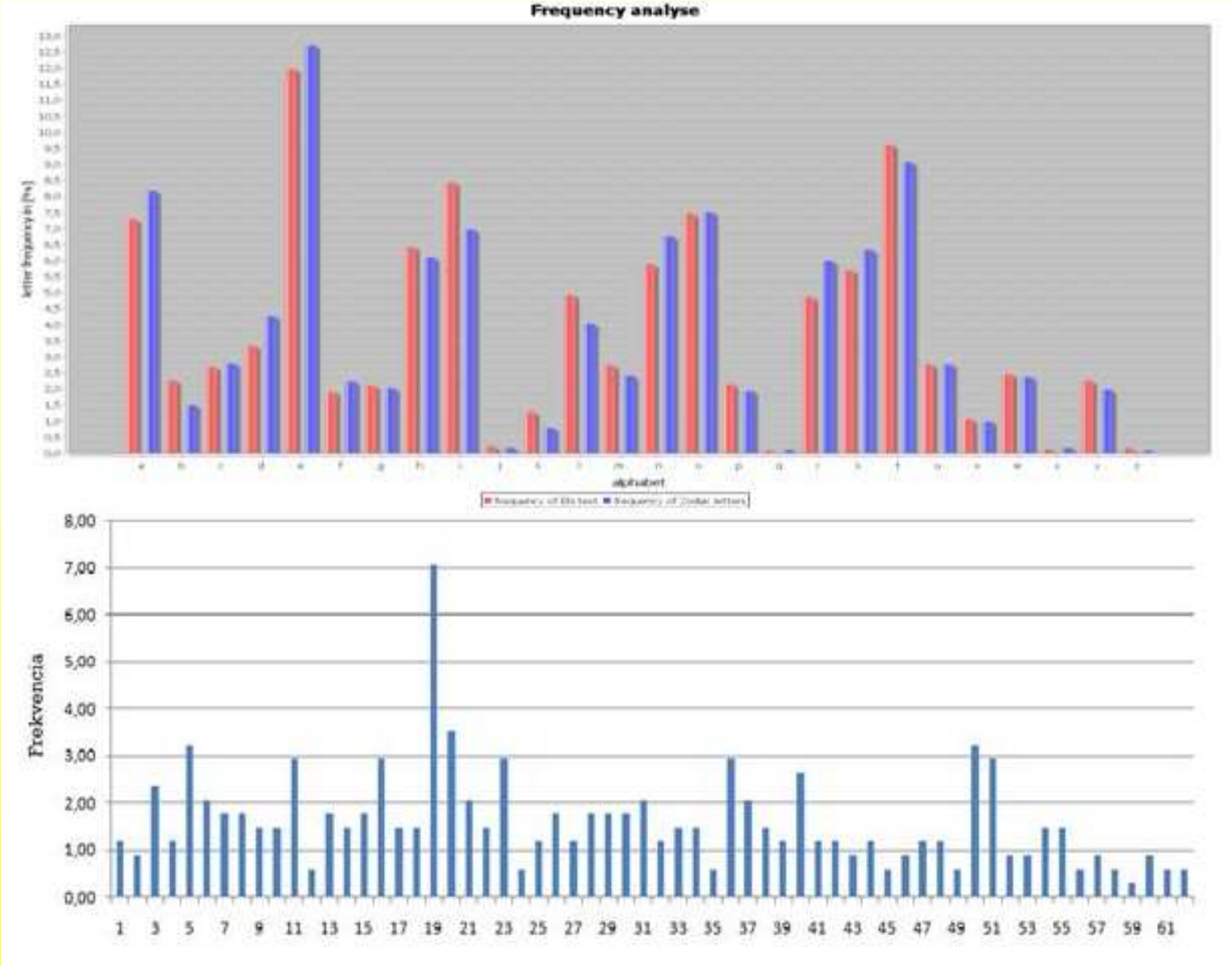
10. - 13.05.2012, ICSAMI 2012 in Malenovice

Introduction

- Substitution cipher - method of encryption
- Units of plaintext are replaced with ciphertext according to regular system
- Units - single letters, pairs, triplets etc.
- Deciphering is performed by inverse operation
- Homophonic cipher - greatly improved classic substitution
- One letter from PT is encrypted by several letters of CT

What do we know?

- Cryptoanalysis of classic substitution is relatively easy - we can use e.g. frequency characteristics of text to guess possible key - pen & paper cryptography
- With introduction of modern computers, this process becomes even more easier (matter of mere seconds)
- These methods are not always successful for classic substitution and almost useless for homophonic ciphers - flattened frequency characteristics
- Main problem of homophonic ciphers is their complexity - vast number of possible combinations for brute-force search
- New methods needed to be introduced


Frequency analyse
alphabet
Frekvencia
8,00
7,00
6,00
5,00
4,00
3,00
2,00
1,00
0,00
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45 47 49 51 53 55 57 59 61

Theoretical solution?

- Stochastic algorithms, heuristic methods - for finding global optimum
- They use fit functions which converge to global optimum (when they are properly set in the beginning)
- Global optimum => maximum value => correct key (in good case)

Motivation

- Primary goal - analyze existing solutions and improve them if possible
- Our research comes out the papers by Diaconis (2009) and Chen & Rosenthal (2012)
- Markov chain Monte Carlo method (Metropolis-Hastings algorithm) is used to effectively solve classic substitution
- Around 10000 iterations needed to find correct key...

Preliminary work

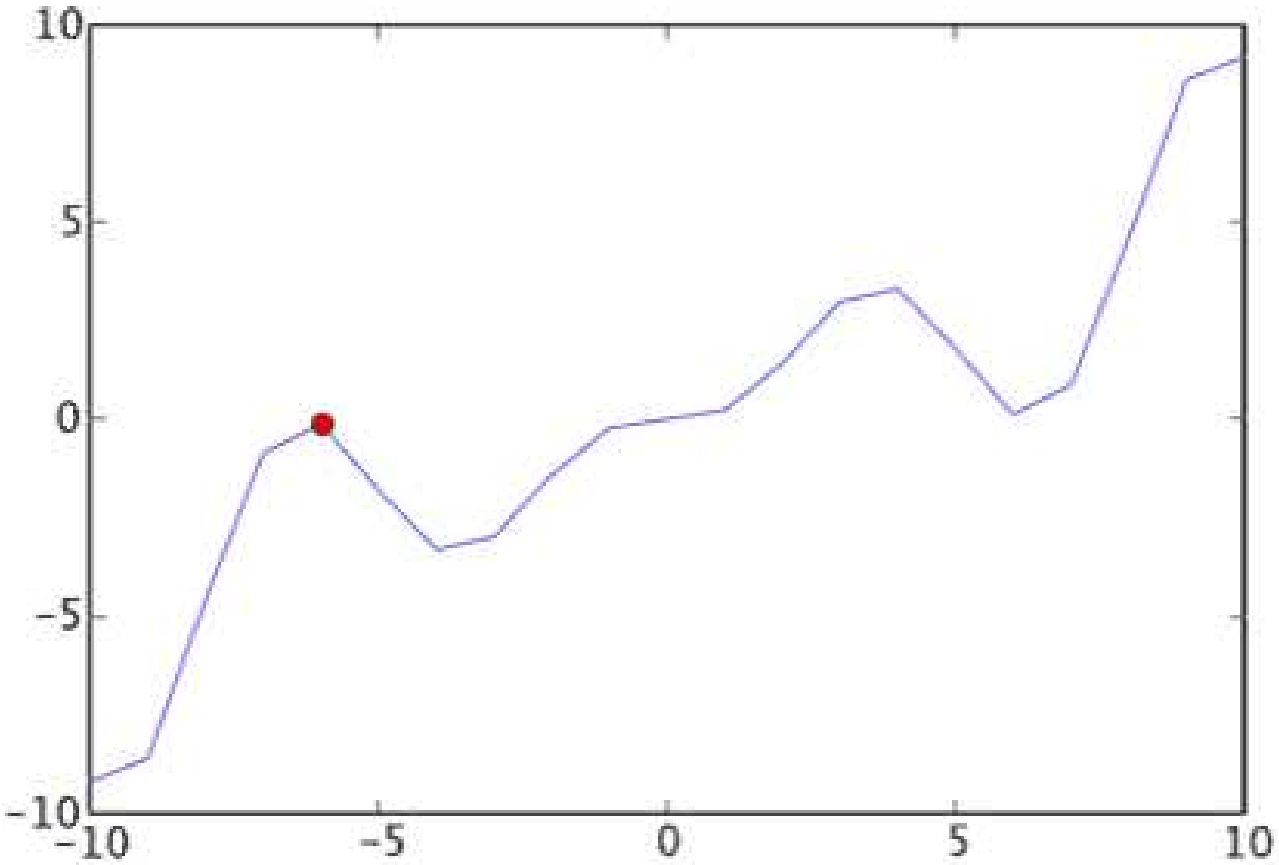
- Thorough examination of some existing algorithms
  - Hill climbing algorithm
  - Metroplis-Hastings algorithm
  - Simulated annealing
- Analysis of possible evaluation functions
- Implementing and testing on simple substitution ciphers
- Quite promising results

Main principle of heuristics

- Based on random search of solution space
- Fit function should converge to global optimum
- Objective - take only better solution
- Getting stuck in local maximas
- Most methods have means to solve this problem:
  - Selecting the best possible fit function
  - Choosing optimal way to get from local maximas
  - Finding optimal method of choosing next state

Hill climbing algorithm

- Basic and easiest of heuristics
- Iterative algorithm
- Random searches solution space and takes only the better solutions according to fit function
- Very effective for solving simple substitutions
- Cornerstone of our research - we designed and tested a few fit functions based on n-gram statistics and dictionary evaluation
- Significantly better results than in literature we based our work on - correct key found in almost all cases
- Main problem of MCMC implementation - also accepts worse solutions with some probability that causes unnecessary increase in iterations needed to solve the cipher
- Our assumption - simple substitution does not have (or insignificant number) local maxima


10
5
0
-5
-10
-10
-5
0
5
10

Switching to homophonic substitution

- Problem #1: too large solution space
- Problem #2: relation between letters of key is not strong enough as in classic substitution
- Problem #3: getting stuck in local maxima

How to avoid getting stuck in local maxima?

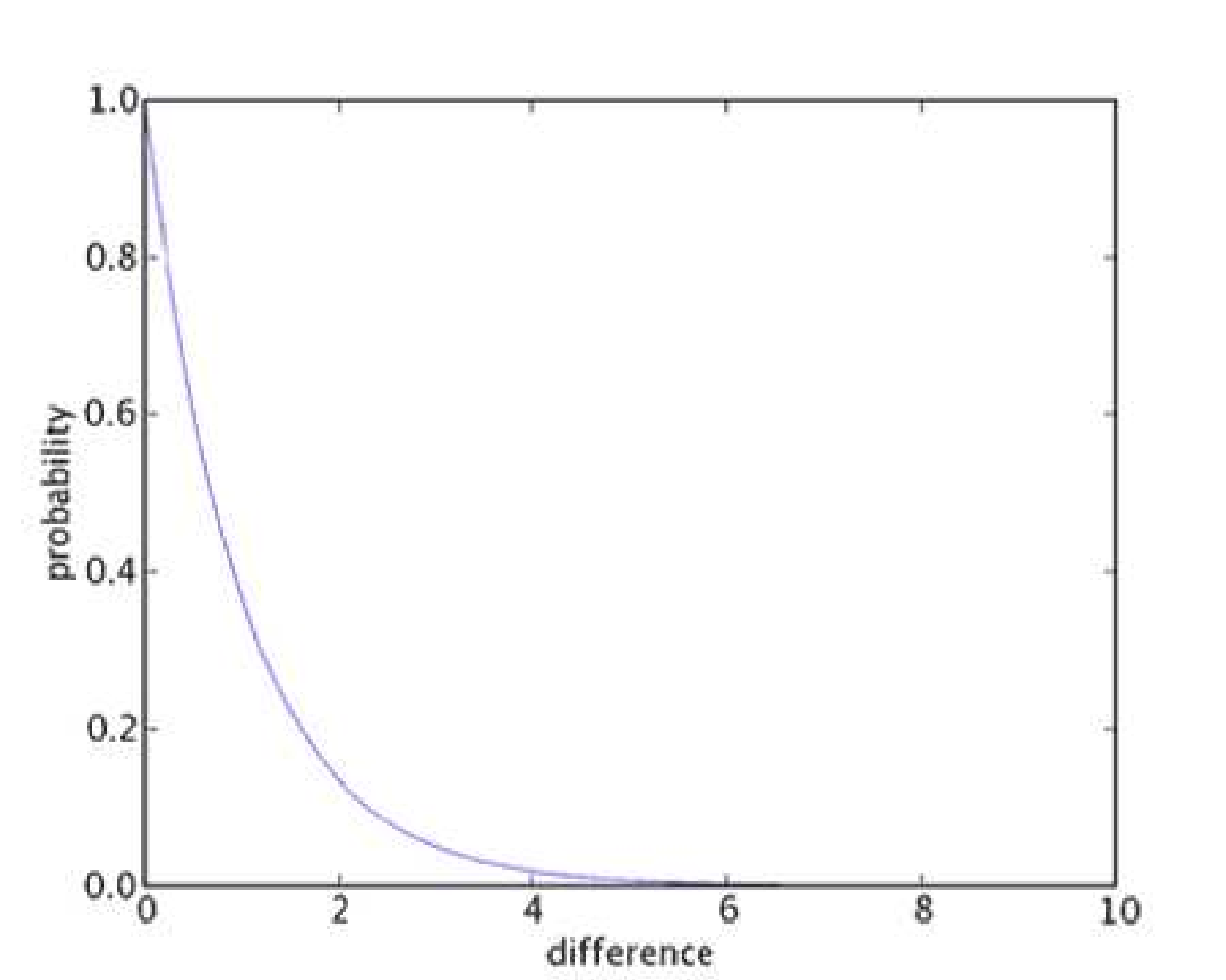
- Make greater changes in the key - can cause omission of correct solution
- Ability to accept worse solutions - we can take 'wrong direction' in solving process

Metroplis-Hastings

- Based on Hill climbing algorithm
- Random selection of initial key
- Change of state by random swap of two letters
- Fit function: $\pi(x) = \prod_{\beta_1,\beta_2} r(\beta_1, \beta_2)^{f_x(\beta_1,\beta_2)}$
- Acceptance of worse solution based on ratio between actual and previous state
- Disadvantage #1: in case of simple substitution - too many unnecessary steps
- Disadvantage #2: ineffective against homophonic ciphers

Simulated annealing

- Probabilistic metaheuristic
- Based on Hill climbing algorithm
- Parametric algorithm derived from physical process of metal annealing
- Probability of accepting worse solution is based on number of iterations
- Possible earlier termination depending on maximum allowed error
- Succesfully solves substitution cipher
- In current state of our implementation ineffective against homophonic substitution


1.0
0.8
0.6
0.4
0.2
0.0
probability
0
2
4
6
8
10
difference

Results so far... part 1

- For simple substitution, the hill climbing method is the most effective
- The best fit functions are absolute bigram and bi-trigram combination difference between possible solution and reference text
- Dictionary-based score function is helpful in later state of searching
- The most effective change of key is to swap two letters - strong dependence in simple substitution
- When inserting whole words - bigger score improvement, convergence rate is very slow, low probability of 'hitting' the correct word position
- Simulated annealing with aforementioned properties is still acceptable but takes more iterations

Results so far... part 2

- Simulated annealing seems to be more effective than Metropolis-Hasting algorithm
- We were able to reduce the number of reqired iterations from 10000 to 1500 in average
- Application capable of effective solving simple substitution
  - Choice of multiple heuristics
  - Choice of fit functions and their combinations
  - Choice of rate how key changes
  - Other parametric setting based on chosen heuristic
- So far ineffective against homophonic ciphers - work still in progress

Future work

- Try out other heuristics, metaheuristics
- Dictionary attacks based on graph algorithms
- Evolution computing - genetic algorithms and swarm intelligence

References

- Chen Jian, Rosenthal Jeffrey S.: *Decrypting classical cipher text using Markov chain Monte Carlo* Statistics and Computing **Volume 22, Issue 2** (March 2012) p. 397-413
- Diaconis Persi: *The Markov Chain Monte Carlo Revolution* Bulletin (New Series) Of The American Mathematical Society **Volume 46, Number 2** (April 2009) p. 179-205

Thank you for your attention!