A F ast Metho d for the Cryptanalysis ofSubstitution CiphersThomas Jak obsen /?yJan uary /8/, /1/9/9/5AbstractIt is p ossible to cryptanalyze simple substitution ciphers /(b oth mono/-and p oly alphab etic/) b y using a fast algorithm based on a pro cess wherean initial k ey guess is re/?ned through a n um b er of iterations/. The algorithm needs to compute thedistribution matrix only once and subsequen t plain text ev aluation isdone b y manipulating this matrix only /, and not b y decrypting theciphertext and reparsing the resulting plain text in ev ery iteration/.The pap er explains the algorithm and it sho ws some of the resultsobtained with an implemen tation in P ascal/.

In a p oly alphab etic cipher more than one suc h mappingis used/.W e assume that the plain text is in English and th us as plain text sym b olsas w ell as ciphertext sym b ols w e use the letters A to Z and space/. Th us amonoalphab etic k ey is a p erm utation of these /2/7 ciphertext sym b ols/. Theciphertext is obtained from the plain text b y replacing eac h sym b ol b ythe corresp onding ciphertext sym b ol in the k ey /. In this w a y w e implicitly geta frequency coun t on w ords b eginning and ending with a particular letter/,namely the /\space/-letter/" and the /\letter/-space/" digram frequencies/./2 A sk etc h of the algorithmThe idea b ehind the algorithm can b e used for ciphertext/-only attac ks onother simple ciphers as w ell/. The algorithm starts b y making an initial guessab out what the k ey is/.

Av erage English has acertain distribution of letters /(space/, E/, T/, A/, O/, N as the most frequen t/, inthat order/) and th us our initial guess is that the most common ciphertextsym b ol is equiv alen t to space/, the second most common is equiv alen t to Eetc/.W e no w need a function/, f /(t /)/, whic h maps a text/, t /, in to a n um b ermeasuring /\ho w close/" the language of t is to the exp ected language of theplain text/, i/.e/./, a function w e can use for describing ho w close w e are to ha vingreco v ered the original plain text/. But /?rst let us in tro duce some notation/.Let m denote the plain text/, c the ciphertext/, and ek
/(t /) the encryptionfunction where k is the k ey used/, so that c /= ek
/(m /)/. Corresp ondingly w eha v e the decryption function dk
/(t /) /= e
/? /1k

/(t /)/.W e de/?ne D /(t /) as the matrix in whic h the elemen ts are the digramfrequencies of a giv en text t /. The ro w headings are the /?rst sym b ols of thedigram and the column headings are the second sym b ols/.E denotes a matrix con taining the exp ected digram frequencies of thelanguage in whic h w e assume the plain text is written/. Digram frequen/-cies for English can b e obtained from sev eral w orks/, including /[/4 /] and /[/7 /]/.W e/, ho w ev er/, compiled our o wn table using /3/0/0/0/0 c haracters of text fromMelville/'s Mob y Dic k /[/6 /]/.

5 10 15 20 25 11.101.45
Number of incorrect
symbols in the key, kf(d(c,k))Figure /1/: f /(d /(c/; k /)/) as a function of an increasingly more inaccurate k ey /.No w w e can state the main parts of the algorithm more formally /.Algorithm /1/:/1/. Construct an initial k ey guess/, k /, based up on the sym b olfrequencies of the exp ected language and the ciphertext/./2/. Let v /= f /(d /(c/; k /)/)/./3 Let k
/0/= k /./4/.

/?nd the corresp onding distribution matrix/. This problem is solv ed in thefollo wing section/./3 A fast approac hJust after step /5 in Algorithm /1 w e ha v ev
/0/= f /(d /(c/; k
/0/)/) /=
Xi/;j
j Dij
/(d /(c/; k
/0/)/) /? Eij
j /=
Xi/;j
j D
/0ij
/(d /(c/; k /)/) /? Eij
j /;where D
/0/(t /) is the distribution matrix/, D /(t /)/, for t with the mo di/?cation thatro ws /? and /? ha v e b een sw app ed after whic h columns /? and /? ha v e b eensw app ed /(or equiv

alen tly sw apping the columns /rst/, then the ro ws/)/.The ab o v e fact can b e used to optimize the algorithm so that w e ha v eto parse the text only once to /nd the distribution matrix/. Construct an initial k ey guess/, k /, based up on the sym b olfrequencies of the exp ected language and the ciphertext/./2/. Let D /= D /(d /(c/; k /)/)/./3/. Let v /=
Pi/;j
j Dij
/? Sw ap t w o elemen ts/, / and / /, in k
/0/./7/.

The strategy used to c ho ose the t w o elemen ts/, / and / /, is describ ed inthe follo wing/. Let s denote the v ector of ciphertext sym b ols rank ed in orderof descending frequency /. Let / /= sa
and / /= sa /+ b
/. Sw ap the sym b ols / and / in k
/0/./6b/. If a /+ b /< /= /2/7 then go to step /7/./6d/. Let a /= b /= /1/.In this w a y the frequencies of the k ey elemen ts /rst sw app ed in k willb e close to eac h other/; /rst w e try to exc hange s/1
/=s/2
/, s/2
/=s/3
/, s/3
/=s/4
/, /./././,s/2/6
/=s/2/7
/, then s/1
/=s/3
/, s/2
/=s/4
/, s/3
/=s/5
/, /./././, s/2/5
/=s/2/7

/, then s/1
/=s/4
/, s/2
/=s/5
/, s/3
/=s/6
/, /././././,s/2/4
/=s/2/7
etc/./, and /nally s/1
/=s/2/7
/- the pair with the largest di/erence regardingsym b ol frequency /./4 P oly alphab etic ciphersF or p oly alphab etic ciphers with n alphab ets the ab o v e algorithm is extendedso that instead of one alphab et w e ha v e sev eral/.

and the columns / and / of the distribution matrix of the previous activ ealphab et/.The ev aluation function b ecomesf /(t /) /=
nXh /=/1
Xi/;j
j D
/(h /)ij
/(t /) /? Eij
j /;and instead of terminating in step /6f if b /= /2/7/, w e let b /= /1/. W e terminatethe algorithm when all the sym b ol pairs of all the alphab ets are exhaustedwithout /nding an y t w o sym b ols to sw ap and at the same time lo w ering theev aluation v alue/./5 Maxim um lik eliho o dIf w e mak e the assumption that the digram frequencies of a language areindep enden t and that eac h one follo ws a Gaussian distribution with mean/ij
/= Eij
and v ariance /
/2ij
for all i /, j /, and if the algorithm do es not getstuc k in lo cal minima /(that is/, if it actually do es minimize the ev aluationfunction/) then w e can /nd an ev aluation function that results in a maxim umlik eliho o d k ey with resp ect to the digram

distribution/.F or a maxim um lik eliho o d attac k our job will b e to maximizeq /=
Yi/;j
/1/?ij
p/2 /?
e
/? /(Dij
/? /?ij
/)
/2/2 /?
/2ij/= r
Yi/;j
e
/?

Ho w ev er/, if the ab o v e men tioned assumptions hold/, then using/(/2/) should result in a more accurate algorithm /(alb eit p erhaps slo w er due tothe extra m ultiplications/./)It should b e men tioned that the general problem of minimizing /(/1/) or/(/2/) giv en D and E can easily b e sho wn to b e at least as hard as the graphisomorphism problem /(consider the incidence matrices of the t w o graphs/)but apparen tly when the plain text is h uman language/, cryptanalysis is quitefeasible/.As ev aluation functions w e also tried to use f /(t /) /=
Pi/;j
qj Dij
/(t /) /? Eij
jand the ob vious candidate f /(t /) /=
Pi/;j
/(Dij
/(t /) /? Eij
/)
/2but it sho w ed up thatgenerally these p erform m uc h w orse than /(/1/)/./6
ResultsFig/. /2 sho ws ho w successful the algorithm w as in attac king texts of v ariouslengths encrypted using b oth mono/- and p oly alphab etic substitution and arandom k ey /.

Length of
ciphertextPercentage of ciphertext
correctly solved
80
70
60
50
40
30
20
1090100
100 200 400 500 600 700 800 900 300 10004 alphabets
5 alphabets1 alphabet
2 alphabets
3 alphabetsFigure /2/: The p ercen tage of text correctly solv ed as a function of the length/.of the ciphertext sym b ols whereas /9/3/% of the text w as correctly resolv ed us/-ing the King/-Bahler algorithm /(according to /[/2 /] and /[/5 /]/)/. It isalso a v ailable in C/./7 ConclusionThe algorithm is quite fast compared to earlier results and it succeeds incryptanalyzing relativ ely small texts though the lengths of solv ed texts arealw a ys m uc h larger than the unicit y distance of the ciphers w e are dealingwith/.As men tioned/, the generalized v ersion of the algorithm /(Algorithm /1/)can b e applied to other simple ciphers although it is probably not p ossibleto optimize it so that one can a v oid reparsing the text in eac h iteration /(lik ein Algorithm /2/)/.

the more mo dern t yp e of encryption algorithms /(IDEA/, DES/, etc/./)Finally it should b e noted/, that there is ro om for other small impro v e/-men ts as w ell/, suc h as using a b etter metho d for deciding whic h pairs tosw ap/. Also w e migh t impro v e the ev aluation function so that it not onlydep ends on digrams/, but p erhaps also on trigrams/, probable w ords and soon/.References/[/1/] Carroll/, J/. M/. and S/. Martin/. /1/9/8/6/. The Automated Cryptanalysis ofSubstitution Ciphers/. Cryptolo gia /. /1/0/(/4/)/: /1/9/3/-/2/0/9/./[/2/] Carroll/, J/. and L/. Robbins/. /1/9/8/7/. The

Automated Cryptanalysis ofP oly alphab etic Ciphers/. /1/1/(/4/)/:
/1/9/3/-/2/0/5/./[/3/] Carroll/, L/. /1/9/9/1/. A lic e in Wonderland /.

and mathematics/.