CHAPTER FOUR

CRYPTANALYSIS OF TRANSPOSITION CIPHER PROBLEMS USING COMBINATORIAL OPTIMIZATION PROBLEMS TECHNIQUES

4.2 Cryptanalysis of TCP

4.2.1 Proposed Cryptanalysis Tools

In general, the literatures assign that, in almost situations, there is no direct solution (decryption) for TCP when using any classical or modern cryptanalysis tools. As usual in cryptanalysis the TCP, the final obtained key, as a result from the cryptanalysis process, will be used to decrypt the cipher text (CT) using decryption key (DK), if the plain text (PT) is not really correct we will still swapping between some wrong positions of the key until we gain good readable text, then we can say that we obtain the actual DK (ADK).

We treat the TCP as a COP. In this manner, we introduce a new study about the diagram (DG), trigram (TG) and quadgram (QG) (4 contagious letters) frequency of PT letters with length L=10000 letters, we called these frequencies as **desired frequencies**. We take in consideration the most frequent samples in PT, so we use the following notations:

D_i^d : Desired Frequency of DG (d-gram) of letter i.
O_i^d : Observed Frequency of DG (d-gram) of letter i.
D_i^t : Desired Frequency of TG (t-gram) of letter i.
O_i^t : Observed Frequency of TG (t-gram) of letter i.
D_i^q : Desired Frequency of QG (q-gram) of letter i.
O_i^q : Observed Frequency of QG (q-gram) of letter i.
Where i='a','b',…,'z'.

In general let $P(X_i^j) = \tilde{X}_i^j$ be the probability of X (=D or =O) desired (D) or observed (O) frequency of j-gram (j=d, t and q), for the letter i s.t.

$$P(X_i^j) = \frac{X_i^j}{L^j} = \tilde{X}_i^j \qquad \ldots(4.1)$$

$$\overline{X}^j = \frac{\sum_{i='a'}^{'z'} X_i^j}{L^j}, \text{ where } j=d,t,q.\ L^d=L-1,\ L^t=L-2 \text{ and } L^q=L-3. \qquad \ldots(4.2)$$

where $\overline{X}^j$ is the arithmetic mean of X_i^j frequency and L is the length of PT.

The results of the desired frequency of the three samples of our study for the English language are illustrated in tables (4.1,2,3) for L=10000.

Table (4.1): Desired frequency of the most common DG (80 DGs).

		Samples (S)													
		1	2	3	4	5	6	7	8	9	10	11	12	13	14
DG	S	AB	AC	AL	AN	AR	AS	AT	BE	CA	CE	CH	CO	DE	DI
	D_i^d	49	41	83	152	67	64	136	47	51	61	46	78	40	46
	S	EA	EC	ED	EE	EI	EL	EM	EN	EO	ER	ES	ET	FT	HA
	D_i^d	95	58	88	44	44	49	58	114	62	160	136	100	41	88
	S	HE	HI	HO	IC	IL	IN	IO	IS	IT	LE	LI	LL	LY	MA
	D_i^d	252	44	49	64	54	210	58	83	113	72	68	54	64	67
	S	ME	NE	NG	NI	NE	NG	NI	NS	NT	OA	OF	OM	ON	OR
	D_i^d	66	49	64	41	49	64	41	48	128	46	105	66	136	82
	S	OT	OU	PE	PR	RA	RE	RM	RO	SA	SE	SI	SO	SS	ST
	D_i^d	52	77	42	78	53	147	60	99	63	96	62	67	54	134
	S	TA	TE	TH	TI	TO	TT	US	UT	VE	WE	---	---	---	---
	D_i^d	69	85	312	161	101	46	50	46	63	60	---	---	---	---

$$\overline{D}^d = \frac{\sum_{i='aa'}^{'zz'} D_i^d}{L^d} \qquad \ldots(4.3)$$

From table (4.1), $\overline{D}^d$=0.6471 and for the most (80) DG frequency, these DG filtered when $\tilde{D}_i^d \geq$TD=0.004.

Table (4.2): Desired frequency of the most common TGs (52 TGs).

		Samples												
		1	2	3	4	5	6	7	8	9	10	11	12	13
TG	S.	ABI	ALL	AND	ARE	ATI	BAB	BIL	COM	CON	ECO	EDI	EIN	EMA
	D_i^t	31	27	66	28	62	31	31	29	23	23	23	22	23
	S.	ENT	ERE	ERS	EST	ETH	FTH	HAT	HEM	HER	ILI	ING	INT	ION
	D_i^t	50	32	28	23	38	35	37	28	44	31	45	42	52
	S.	IST	ITI	ITY	LIT	MAT	NCE	NTH	OBA	OFT	OME	OTH	OUT	PRO
	D_i^t	35	28	28	35	29	35	41	31	34	25	27	24	53
	S.	REA	ROB	SOF	STA	STH	STO	THA	THE	THI	TIC	TIO	TIS	TTH
	D_i^t	24	37	34	22	33	25	38	219	23	23	45	22	27

$$\overline{D}^t = \frac{\sum_{i='aaa'}^{'zzz'} D_i^t}{L^t} \qquad \ldots(4.4)$$

From table (4.2), $\overline{D}^t$ = 0.1901, for the most (52) TG frequency, these filtered when $\tilde{D}_i^t \geq TT$ =0.0022 is a TG-threshold.

Table (4.3): Desired frequency of the most common QGs (21 QGs).

		Samples										
		1	2	3	4	5	6	7	8	9	10	11
QG	S.	ABIL	ATIO	BABI	BILI	ETHE	FTHE	ILIT	INTH	LITY	NTHE	OBAB
	D_i^q	31	26	31	31	28	33	31	22	24	27	31
	S.	OFTH	OTHE	PROB	ROBA	STHE	THAT	THEM	THEO	THER	TION	---
	D_i^q	30	22	37	31	23	29	28	20	34	45	---

$$\overline{D}^q = \frac{\sum_{i='aaaa'}^{'zzzz'} D_i^q}{L^q} \qquad \ldots(4.5)$$

From table (4.3), $\overline{D}^q$ = 0.0614, for the most (21) QG frequency, these filtered when $\tilde{D}_i^q \geq TQ$ =0.002 is a QG-threshold.

4.2.2 Proposed Cryptanalysis Objective Functions

From equations (4.3-4.5), the **sum of the most high frequency** (SMHF) for PT and CT are calculated in equations (4.6) and (4.7) respectively.

$$SMHF(M) = \overline{D}^{d} + \overline{D}^{t} + \overline{D}^{q} \quad \ldots(4.6)$$

$$SMHF(C) = \overline{O}^{d} + \overline{O}^{t} + \overline{O}^{q} \quad \ldots(4.7)$$

The $\overline{O}^{j}$ of letters of CT are the corresponding to the letters of PT frequency $\overline{D}^{j}$ mentioned in equation (4.6), where j=d, t, q.

It's clear that for L=10000 PT letters the SMHF(M)=0.6471+0.1901 +0.0614=0.8986.

Of course we are interest in SMHF for PT and CT. Table (4.4) shows the values of SMHF for PT and CT for text with length L=(start=1000,(step=1000),end=10000) letters.

Table (4.4): Values of SMHF(M) and SMHF(C) for text with different L.

SMHF	L ×1000 letters										Av.
	1	2	3	4	5	6	7	8	9	10	
SMHF(M)	1.375	0.979	0.929	0.894	0.868	0.862	0.887	0.923	0.930	0.899	0.955
SMHF(C)	0.848	0.587	0.549	0.527	0.512	0.498	0.476	0.485	0.488	0.491	0.546

Now we have another measure to diagnose the TC which is called the **coincidence of desired frequency** (CDF) for PT. This value can be calculated as follows:

$$CDF = \sum_{i='aa'}^{'zz'} \left| \tilde{D}_i^{d} - \tilde{O}_i^{d} \right| + \sum_{i='aaa'}^{'zzz'} \left| \tilde{D}_i^{t} - \tilde{O}_i^{t} \right| + \sum_{i='aaaa'}^{'zzzz'} \left| \tilde{D}_i^{q} - \tilde{O}_i^{q} \right| \quad \ldots(4.8)$$

The frequency in equation (4.8), are for all combinations of two, three and four letters sample in PT or CT. Table (4.5) shows the CDF values for different L for PT and CT.

Table (4.5): The CDF values for different L for PT and CT.

CDF	L ×1000 letters										Av.
	1	2	3	4	5	6	7	8	9	10	
CDF(M)	0.352	0.225	0.206	0.176	0.159	0.117	0.080	0.044	0.030	0.000	0.139
CDF(C)	0.627	0.593	0.600	0.590	0.593	0.594	0.594	0.597	0.597	0.598	0.598

Note that the CDF(M) values, for different text lengths, is in continuous decreasing, while the CDF(C) values are in stable context. Notes that the worst case

for CDF(M) function is 0.352 for L=1000 and worst case for CDF(C) function is 0.59 for L=4000 (shaded cells).

4.2.3 TCP Formulation

The cryptanalysis of TCP can be treated as a COP with two objective functions (TOF). Suppose that the encryption key (EK) treated as a sequence $\sigma=(1,2,\ldots,n)$, where n is the length of the EK (or DK), s.t.

$$\text{TOF}(n,\sigma)=\{\text{Max SMHF, Min CDF}\} \qquad \ldots(4.9)$$

The two objectives aggregated into one composite single objective function for n key length for the sequence σ (SOF(n,σ)), where SOF is a single objective function. Since $0\le\text{CDF}\le 1$ and $\text{Min}\{\text{CDF}\}\equiv \text{Max}\{1-\text{CDF}\}$, hence the TCP can be written as follows:

$$\left.\begin{array}{l} \text{SOF}(n,\sigma)=\text{Max}\ \{\text{SMHF}+1-\text{CDF}\} \\ \text{s. t.} \\ 0.8622 \le \overline{O}^{d}+\overline{O}^{t}+\overline{O}^{q}\le 1.375 \\ 0\le \sum_{i='aa'}^{'zz'}\left|\tilde{D}_i^{d}-\tilde{O}_i^{d}\right|+\sum_{i='aaa'}^{'zzz'}\left|\tilde{D}_i^{t}-\tilde{O}_i^{t}\right|+\sum_{i='aaaa'}^{'zzzz'}\left|\tilde{D}_i^{q}-\tilde{O}_i^{q}\right|\le 0.352 \\ \overline{O}^{d},\overline{O}^{t},\overline{O}^{q},\tilde{O}_i^{d},\tilde{O}_i^{t},\tilde{O}_i^{q}\ge 0 \end{array}\right\} \quad \ldots(P)$$

Where $\tilde{D}_i^{d},\tilde{D}_i^{t},\tilde{D}_i^{q}$ are known from tables (4.1), (4.2) and (4.3).