CARDIFF UNIVERSITY
EXAMINATION PAPER

Academic Year: 2001/2002
Examination Period: SPRING
Examination Paper Number: CM0321
Examination Paper Title: Data Security
Duration: Two hours

Do not turn this page over until instructed to do so by the Senior Invigilator.

Structure of Examination Paper:

There are THREE pages.
There are FOUR questions in total.
There are no appendices.

The maximum mark for the examination paper is 60 marks = 100% (the equivalent of three complete questions) and the mark obtainable for a question or part of a question is shown in brackets alongside the question.

Students to be provided with:

The following items of stationery are to be provided:
ONE answer book.

Instructions to Students:
Answer THREE questions.
1. A multi-user computer system with a commercial operating system and large number of on-line users is being installed by a company that requires a moderate level of security. You are asked to advise on various aspects of security.

(a) What general advice would you give?  

(b) What suggestions would you make aimed at preventing unauthorised access to the system?  

(c) What general procedures would you suggest to reduce the threat from (a) Viruses, (b) Trojan Horses?  

(d) There is a telecommunications data connection to another site which is used to transfer large amounts of commercially confidential data. How technically difficult would it be for an unauthorised person to monitor such a link? What type of encryption might be suitable in this case?  

2. (a) Explain what is meant by a binary Hamming code of length $2^r - 1$.  

(b) Explain the simplest decoding rule for Hamming codes.  

(c) The parity check matrix of a Hamming code is

$$H = \begin{pmatrix} 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 \end{pmatrix}$$

[Warning: $H$ is not in standard form.]

(i) What is the standard form Generator matrix of this code, how many codewords are there in the code, and what is the minimum distance of the code?  

(ii) The vector $\{0, 1, 1, 1, 1, 0\}$ is received. What is its syndrome with respect to $H$? Hence decode the word correctly using (b).
3. (a) Describe Shamir’s method for sharing a secret key among \( n \) persons in such a way that any subset of \( t \) or more can recover the key but no subset of less than \( t \) can do so. [5]

(b) In the following case \( n = 5 \) and \( t = 4 \) and Shamir’s method has been used to provide information about the secret key. Persons 1, 2, 4 and 5 decide to recover the key. Their numbers are

\[
f(1) = 2, f(2) = 7, f(4) = 8, f(5) = 3 \text{ and all arithmetic is (mod 19).}
\]

Find the polynomial and hence the key and also the value of \( f(3) \). [8]

(c) The prime used in Shamir’s method is about \( p \approx 10^{200} \), \( t = 20 \), and the secret numbers are uniformly randomly selected between 1 and \( p - 1 \). The secret key is now chosen to be the concatenation of the coefficients of the polynomial. If the encryption system is known and the government can check \( 10^{10} \) secret keys per second, given an encrypted message how many years would it take to recover the plain text? Is this feasible? [Assume that exhaustive search is the only approach. There are approximately \( 3.15 \times 10^7 \) seconds per year.] [7]

4 (a) Describe Euclid's algorithm for determining the highest common factor \( \text{hcf}[l, m] \) of positive integers \( l, m \) and explain why it works. [4]

(b) Explain how Euclid's algorithm can be used to find one solution of \( ax + by = d \) in integers \( x, y \), where \( a, b, d \) are integers and \( \text{hcf}[a, b] \) divides \( d \). [4]

(c) Explain the RSA algorithm for public key encryption/decryption and the relevance of (b). [6]

(d) A public key system which uses the RSA algorithm is based on the modulus \( n = 517 \) and the (private) decryption key \( d = 307 \).

(i) What is the encryption key? [3]

(ii) What is the encrypted form of the number 245? [3]
Solutions

1 [All early bookwork - meant to be straightforward question to settle everyone down]

(i) What general advice would you give? [5]

1. Keep back-up copies of all important programs and data. Do this on a strict rotation basis with a long time period. Keep duplicates in a geographically distinct secure location.

2. Never allow free-gifts or downloaded software onto the system without thorough prior checking.

3. Write protect any data storage medium that is intended only for reading.

4. Keep passwords secret, don’t write them down, and change them regularly (a high percentage of users never change their passwords).


6. When users leave the company remove their names and passwords from the list of authorised users.

7. Do not leave a logged-on terminal unattended.

[5 marks for any 5 points]

(ii) What suggestions would you make aimed at preventing unauthorised access to the system? [5]

Not so easy because authorised users are often very careless with their passwords. However, the following measures can improve both denial and detection.

1. User ID backed up by one-way enciphered passwords. Passwords should never be echoed back or transmitted in clear.

2. Automatic rejection of users after (say) three incorrect submissions of passwords. The account should be frozen and the security officer or system administrator be informed.

3. If not (2) then impose a time delay of several seconds before allowing a re-try on a password. (This prevents a computer being used to try large numbers of passwords.)

4. Authentication of signatures by public key encryption.

5. Automatic logging of all user interaction.

(iii) What general procedures would you suggest to reduce the threat from (a) Viruses, (b) Trojan Horses? [5]

1. It is possible for some specific viruses to "inoculate" a file by causing the virus to "think" that it is already present.

2. A simple regular test for the presence of a virus or other parasite is to see if the length of any file has increased.

3. Include a check-sum, using an unpublished one-way encryption algorithm, as an appendix to each block of program or data. Any modification will then be detected by failure of the check sum. This may involve a customised operating system, or customised software - so is often not practical in a commercial environment.
4. Detect Trojan horses (etc) by a dynamic trace program which reveals sections of unused code, thus helping to locate the alien code.

5. Encrypt the programs and data.

6. Install an email virus checker.

(iv) How technically difficult would it be for an unauthorised person to monitor such a link? What type of encryption might be suitable in this case?

This is relatively easy for electrical communication cables and somewhat more difficult with fibre-optical lines. It is quite hard to detect.

With both kinds of links, constant monitoring of the link state with auto-logging of link-interruptions or variations of link quality plus identification of the location of any interruption can be technically feasible - but of course expensive.

Therefore all data sent along such links should be encrypted, with a level of sophistication proportionate to the value/secrecy of the data.

In this particular case the link is high bandwidth so the encrypt/decrypt algorithm must be implemented in hardware. The level of security is commercial so the DES would be appropriate (if not particularly secure these days).
2. (a) [Book work] Explain what is meant by a binary Hamming code of length $2^r - 1$.

**Definition.** Hamming code. Let $r$ be a positive integer and $n = 2^r - 1$. Let $H$ be an $r \times n = r \times (2^r - 1)$ matrix whose columns are all the distinct non-null vectors of $V(n, 2)$ (the vector space in $n$-dimensions over the field $\mathbb{F}_2$). Then the code having $H$ as its parity check matrix is called a binary Hamming code and is denoted by Ham($r, 2$).

**Note:** $H$ is the parity check matrix of a binary $[n, k]$ code where $n = 2^r - 1$ and $k = n - r$.

(b) [Book work] Explain the simplest decoding rule for Hamming codes.

- **If** the syndrome $Hy^T$ is non-zero it must be some column of $H$ (because $H$ contains all non-null columns), suppose it is the $j$th column of $H$ then correct the $j$th bit of $y$. **Else** $y$ is assumed to be correct.

(c) The parity check matrix of a Hamming code is

$$H = \begin{pmatrix}
0 & 0 & 0 & 1 & 1 & 1 & 1 \\
0 & 1 & 1 & 0 & 0 & 1 & 1 \\
1 & 0 & 1 & 0 & 1 & 0 & 1
\end{pmatrix}$$

[Warning: $H$ is not in standard form.]

(i) **What is the standard form Generator matrix of this code?**

Note here $r = 3$, $n = 7$ and $k = 4$. Rearranging $H$ to be in standard form $stdH = [-A^T, I_{n-k}]$ we have

$$stdH = \begin{pmatrix}
0 & 1 & 1 & 1 & 1 & 0 & 0 \\
1 & 0 & 1 & 1 & 0 & 1 & 0 \\
1 & 1 & 0 & 1 & 0 & 0 & 1
\end{pmatrix}$$

Hence the generator matrix $stdG = [I_r, A]$ is

$$stdG = \begin{pmatrix}
1 & 0 & 0 & 0 & 0 & 1 & 1 \\
0 & 1 & 0 & 0 & 1 & 0 & 1 \\
0 & 0 & 1 & 0 & 1 & 1 & 0 \\
0 & 0 & 0 & 1 & 1 & 1 & 1
\end{pmatrix}$$

How many codewords are there in the code

There are $q^k = 2^4 = 16$ codewords.

What is the minimum distance of the code?

Theory tells us that the minimum distance of the code is 3.

(ii) The vector $[0, 1, 1, 1, 1, 1, 0]$ is received. What is its syndrome? Hence decode the word correctly.

The syndrome is $H[0, 1, 1, 1, 1, 1, 0]^T = [1, 1, 0]$. This is the sixth column of $H$ (N.B. Not $stdH$) hence there is an error in the sixth bit and the corrected form of the received word is $[0, 1, 1, 1, 1, 0, 0]$. 


3. (a) [Bookwork] Describe Shamir’s method for sharing a secret key among \( n \) persons in such a way that any subset of \( t \) or more can recover the key but no subset of less than \( t \) can do so.

Shamir’s method. The trusted administrator chooses a polynomial of degree \( t - 1 \) with integer coefficients (mod \( p \)), where \( p \) is a large prime \( p >> n \). The constant term in this polynomial is \( K \) (the secret number), the other coefficients are chosen at random (mod \( p \)).

Let this polynomial be

\[
f(x) = K + a_1x + a_2x^2 + \ldots + a_{t-1}x^{t-1}
\]

The \( n \) recipients are allocated the identifying labels 1, 2, ..., \( n \) and the \( i \)th recipient is told the value of \( f(i) \) (mod \( p \)). In order to recover the secret number \( K \) it is sufficient to find the polynomial \( f(x) \), because then \( K = f(0) \).

- This can be done if and only if \( t \) pairs of values \((i, f(i))\) are available.

(b) In the following case \( n = 5 \) and \( t = 4 \) and Shamir’s method has been used to provide information about the secret key. Persons 1, 2, 4 and 5 decide to recover the key. Their numbers are

\[
\begin{align*}
f(1) &= 2, \quad f(2) = 7, \quad f(4) = 8, \quad f(5) = 3 \quad \text{and all arithmetic is (mod 19)}.
\end{align*}
\]

Find the polynomial and hence the key and also the value of \( f(3) \).

From the question

\[
\begin{align*}
f(1) &= K + a_1 + a_2 + a_3 \equiv 2 \pmod{19} \\
f(2) &= K + 2a_1 + 4a_2 + 8a_3 \equiv 7 \pmod{19} \\
f(4) &= K + 4a_1 + 16a_2 + 7a_3 \equiv 8 \pmod{19} \\
f(5) &= K + 5a_1 + 25a_2 + 11a_3 \equiv 3 \pmod{19}
\end{align*}
\]

Solving these congruences by Gaussian elimination gives \( \{K, a_1, a_2, a_3\} = \{1, 2, 7, 11\} \). Therefore the secret key is \( K = 1 \) and \( f(3) = 6 \pmod{19} \).

(c) The prime used in Shamir’s method is about \( p = 10^{200} \), \( t = 20 \), and the secret numbers are uniformly randomly selected between 1 and \( p-1 \). The secret key is now chosen to be the concatenation of the coefficients of the polynomial. If the encryption system is known and the government can check \( 10^{10} \) secret keys per sec, given an encrypted message how long in years would it take to recover the plain text?

[There are approximately \( 3.15 \times 10^7 \) sec per year.]

If \( p = 10^{200} \) then the expected size of a polynomial coefficient is about half this, i.e \( 5 \times 10^{199} \), i.e. a decimal number about 200 digits long. (N.B. Not 100 digits long). A concatenation of \( t = 20 \) such numbers therefore has expected length about 4000 decimal digits. Thus if the government can check \( 10^{10} \) such numbers per sec. It will take of the order of

\[
\frac{10^{4000}}{10^{10}} = 10^{3990} \text{ secs}
\]

\[
= \frac{10^{3990}}{3.15 \times 10^7} = 3.17 \times 10^{3982} \text{ years}
\]

Is this feasible? No. (Anyone who says ‘yes’ ought to fail automatically!)
4.(a) [Bookwork] Describe Euclid’s algorithm for determining the highest common factor $\text{hcf}[l, m]$ of two integers $l, m$. [4]

The algorithm proceeds as follows:

We can assume that $x > y$. Put $x_0 = x$, $x_1 = y$. Then compute $x_2$, $x_3$, ... defined by
\[ x_{n+1} = m_n x_n + x_{n+1} \quad (0 \leq x_{n+1} < x_n) \]
If $x_{n+1} = 0$ then $x_n = \text{hcf}[x, y]$ (so stop). If $x_{n+1} \neq 0$ then increase $n$ by one and repeat the process.

The algorithm works because: (1) $\text{hcf}[x, y]$ coincides with $\text{hcf}[r, x]$, where $r$ is the remainder of $y$ after division by $x$ and (2) the recursion eventually falls into the $\text{hcf}[., 0]$ case because the remainder $r < x$ and so the remainders form a decreasing sequence of non-negative integers (and it follows from the Well Ordering Principle that such a set has zero as its minimum).

(b) [Bookwork] Explain how Euclid’s algorithm can be used to solve $ax + by = d$ in integers $x$, $y$, where $a$, $b$, $d$ are integers and $\text{hcf}[a, b]$ divides $d$. [4]

If $\text{hcf}[a, b]$ divides $d$ to solve this equation in integers $x$ and $y$ we execute Euclid’s algorithm to find $\text{hcf}[a, b]$. This gives a set of equations
\[ x_0 = a, \quad x_1 = b \]
\[ x_0 = m_1 x_1 + x_2 \quad (0 < x_2 < x_1) \]
\[ x_{n-1} = m_n x_n + d \]
\[ x_n = m_n x_d + 0 \quad (\text{i.e. } d = x_{n-1}) \]

Next substitute for $x_{n-1}$ in terms of $x_{n-2}, x_{n-3}, ...$ etc. until we finally reach an identity involving $x_0 (= a)$ and $x_1 (= b)$ of the type $d = ux_0 + vx_1$, i.e. $d = au + bv$.

(c) [Bookwork] Explain the RSA algorithm for public key encryption/decryption and the relevance of (b). [6]

The system is based on choosing three integers $(e, p, q)$ where $p$ and $q$ are large primes $(\approx 10^{100})$ and $e$ is coprime (has no factors in common) to $(p - 1)(q - 1)$.

Let $n = pq$. Then $e$ and $n$ are made public and only the owner of the system knows $p$ and $q$.

To encrypt the message:

1. Convert the message into a stream of decimal digits (the method for doing this is not secret, indeed it must be known).
2. Cut the stream into blocks of equal length and convert these to decimal numbers $m_1, m_2, ..., $ where $m_i < n$.
3. Raise each number $m_i$ to the power $e$ and replace $m_i^e$ by its remainder $r_i \pmod{n}$ (so $m_i^e = r_i \pmod{n}$ in mathematical notation).
4. The encrypted message is then the stream $r_1r_2r_3...$

To decrypt the message:

1. Let $d$ be the number that has the property
\[ ed \equiv 1 \pmod{(p-1)(q-1)} \]
(this is not too hard to compute, it can be done by using **Euclid’s algorithm**, and can be worked out once and for all and saved).

2. Then

\[ r_i^d \equiv (m_i^e)^d \equiv m_i^{ed} \equiv m_i^{(p\cdot q) - 1} \equiv m_i \pmod{n} \]

by the Euler-Fermat theorem, and the decrypted message is \(m_1, m_2, m_3,...\)

(d) A public key system which uses the RSA algorithm is based on the modulus \(n = 517\) and the (private) decryption key \(d = 307\).

(i) **What is the encryption key?**

\[ pq = 517 \text{ factorises to } p = 47 \text{ and } q = 11. \text{ Thus } \varphi(pq) = (p-1)(q-1) = 46 \cdot 10 = 460. \]

\(d = 307\) so we have to solve \(307e \equiv 1 \pmod{460}\). This can be done with Euclid's algorithm (but the numbers have been chosen so that it is easy to find a solution by inspection) giving \(e = 3\), which is the public encryption key.

(ii) **What is the encrypted form of the number 245?**

Now we know \(e\) the encrypted form of 245 is \(245^3 \equiv 60 \pmod{517}\), i.e. 60.