Lecture 6: Who owns ideas?

6.1 What is "intellectual property"?

The term "intellectual property" can refer to ideas protected by copyright, patent, trademark, or trade secret. *Intellectual Property and the National Information Infrastructure: The Report of the Working Group on Intellectual Property Rights* (November, 1995) available through the web page `http://www.uspto.gov/web/offices/com/doc/ipnii/` discusses all of these. I'll write only about the first of these, copyright, although digital information could easily infringe on ideas protected by the other classical restrictions.

Copyright

Copyright may be different in different countries. A webpage of the U.S. government, `http://www.loc.gov/copyright/circs/circ1.html`, has this description of copyright. The "..." indicates deletion of some technical phrases.

> Copyright is a form of protection provided by the laws of the United States ... to the authors of original works ... including literary, dramatic, musical, artistic, and certain other intellectual works. This protection is available to both published and unpublished works. ... the 1976 Copyright Act generally gives the owner of copyright the exclusive right to do and to authorize others to do the following:
>
> - To reproduce the work in copies or phonorecords;
> - To prepare derivative works based upon the work;
> - To distribute copies or phonorecords of the work to the public by sale or other transfer of ownership, or by rental, lease, or lending;
> - To perform the work publicly, in the case of literary, musical, dramatic, and choreographic works, pantomimes, and motion pictures and other audiovisual works;
> - To display the copyrighted work publicly, in the case of literary, musical, dramatic, and choreographic works, pantomimes, and pictorial, graphic, or sculptural works, including the individual images of a motion picture or other audiovisual work; and
> - In the case of sound recordings, to perform the work publicly by means of a digital audio transmission.
>
> In addition, certain authors of works of visual art have the rights of attribution and integrity ...
>
> It is illegal for anyone to violate any of the rights provided by the copyright law to the owner of copyright. These rights, however, are not unlimited in scope. Sections ... of the 1976 Copyright Act establish limitations on these rights. In some cases, these limitations are specified exemptions from copyright liability. One major limitation is the doctrine of "fair use" .... . In other instances, the limitation takes the form of a "compulsory license" under which certain limited uses of copyrighted works are permitted upon payment of specified royalties and compliance with statutory conditions.

The Digital Dilemma

A recent book-length report on digital intellectual property was published by the (U.S.) National Academy of Sciences. It is *The Digital Dilemma: Intellectual Property in the Information Age* and is available at `http://www.nap.edu/books/0309064996/html/`. Here's the beginning of the Executive Summary:

> THE ORIGINS OF THE DIGITAL DILEMMA
>
> Borrowing a book from a local public library would seem to be one of the most routine, familiar, and uncomplicated acts in modern civic life. A world of information is available with little effort and almost no out-of-pocket cost. Such access to information has played a central role in American education and civic life from the time of Thomas Jefferson, who believed in the crucial role that knowledge and an educated populace play in making democracy work. Yet the very possibility of borrowing a book, whether from a library or a friend, depends on a number of subtle, surprisingly complex, and at times conflicting elements of law, public policy, economics, and technology, elements that are in relative balance today but may well be thrown completely out of balance by the accelerating transformation of information into digital form.
>
> The problem is illustrated simply enough: A printed book can be accessed by one or perhaps two people at once, people who must, of course, be in the same place as the book. But make that same text available in electronic form, and there is almost no technological limit to the number of people who can access it simultaneously, from literally anywhere on the planet where there is a telephone (and hence an Internet connection).
>
> At first glance, this is wonderful news for the consumer and for society: The electronic holdings of libraries (and friends) around the world can become available from a home computer, 24 hours a day, year-round; they are never "checked out". These same advances in technology create new opportunities and markets for publishers.
>
> But there is also a more troublesome side. For publishers and authors, the question is, How many copies of the work will be sold (or licensed) if networks make possible planet-wide access? Their nightmare is that the number is one. How many books (or movies, photographs, or musical pieces) will be created and published online if the entire market can be extinguished by the sale of the first electronic copy?
>
> The nightmare of consumers is that the attempt to preserve the marketplaces leads to technical and legal protections that sharply reduce access to society's intellectual and cultural heritage, the resource that Jefferson saw as crucial to democracy. . . .

The page `http://www.arl.org/info/frn/copy/primer.html` discusses the DMCA, a relatively new law. The act extends certain copyrights (Mickey Mouse's masters are *very* happy about this!) and also changes some U.S. laws and practices to make them agree more with worldwide standards. U.S. laws on intellectual property can be seen through the web page `http://fedlaw.gsa.gov/fedfra23.htm`. A general reference on intellectual property is `http://www.ip-surveys.com/links/links.html` which includes links to "world" practices, not just the U.S.

6.2 Protecting intellectual property in the digital era

Encryption

Encryption methods used to protect intellectual property may rely on both "software" or "hardware". Some software methods relying on the difficulty of certain algorithms to be unraveled without knowledge of their keys. Special purpose hardware has been designed to forestall copying or access to intellectual property. This could take the form of special computer chips encased in decoder boxes or special attachments for computers ("dongles"). These can be unwieldy and unpopular. Digital property when converted to plaintext can perhaps just be copied unless additional intricate arrangements are made. Also, decryption can be elaborate and take time, and its intricacies can be another level of complexity subject to malfunction during delivery of the property. Additionally, of course, encryption schemes can be broken. A notorious recent case is the cracking of the DVD protection scheme.

Here's a paragraph from an official Memorandum of the U.S. District Court published in January 2000 (source: a web page at `www.eff.org`):

> DVDs contain motion pictures in digital form, which presents an enhanced risk of unauthorized reproduction and distribution because digital copies made from DVDs do not degrade from generation to generation. Concerned about this risk, motion picture companies, including plaintiffs, insisted upon the development of an access control and copy prevention system to inhibit the unauthorized reproduction and distribution of motion pictures before they released films in the DVD format. The means now in use, Content Scramble System or CSS, is an encryption-based security and authentication system that requires the use of appropriately configured hardware such as a DVD player or a computer DVD drive to decrypt, unscramble and play back, but not copy, motion pictures on DVDs. CSS has been licensed to hundreds of DVD player manufacturers and DVD content distributors in the United States and around the world.

The licensees are required to keep the algorithms and keys secret, and to install tamper-proof hardware in each player which would contain the keys needed to read those DVD's legal for the player's area in the world. The DVD organization divided the world into several distinct areas: discs and players could each work only in designated areas.

CSS turns out to be a rather simple encryption scheme relying on 40 bit keys (not adequate by current encryption standards). The system was reverse engineered: that is, observation of the inputs and outputs of the system were used to figure out how the system worked. Also, one of the manufacturers mistakenly allowed some of the keys to be seen. Even extremely secure cryptosystems may be vulnerable when users make mistakes.

The following quotes are taken from pages on `http://www.wired.com`. The first is from, of course, a representative of those wishing to protect their intellectual property. The second is from the "crackers".

> "The circulation through the Internet of the illegal and inappropriate software is against the stream of copyright protection. Toshiba, which has led the establishment of the DVD format and is the chair-company of the DVD Forum, feels it is a great pity," wrote Masaki Mikura, manager of the strategic partnership and licensing division at Toshiba Ltd.
>
> Johansen and his partners were able to guess more than 170 working keys by trial and error before finally just giving up to go do something else. "I wonder how much they paid for someone to actually develop that weak algorithm," said Johansen. "It's a very weak encryption algorithm."

Steganography

From the web page `http://www.cl.cam.ac.uk/~fapp2/steganography/`, which is called "the information hiding homepage digital watermarking & steganography":

> While cryptography is about protecting the content of messages ... **steganography** is about concealing their very existence. It comes from Greek roots and literally means "covered writing," and is usually interpreted to mean hiding information in other information.
>
> Until recently, information hiding techniques received very much less attention from the research community and from industry than cryptography, but this is changing rapidly

> ... The main driving force is concern over protecting copyright; as audio, video and other works become available in digital form, it may be that the ease with which perfect copies can be made will lead to large-scale unauthorized copying which will undermine the music, film, book and software publishing industries. There has therefore been significant recent research into "watermarking" (hidden copyright messages) and "fingerprinting" (hidden serial numbers or a set of characteristics that tend to distinguish an object from other similar objects); the idea is that the latter can be used to detect copyright violators and the former to prosecute them.
>
> ©1997-1999 by Fabien A. P. Petitcolas, Computer Laboratory, University of Cambridge

From the webpage `http://members.tripod.com/steganography/stego.html`:

> **What is Steganography?**
>
> In an ideal world we would all be able to openly send encrypted mail or files to each other with no fear of reprisals. However there are often cases when this is not possible, either because you are working for a company that does not allow encrypted email or perhaps the local government does not approve of encrypted communication (a reality in some parts of the world). This is where steganography can come into play.
>
> Steganography simply takes one piece of information and hides it within another. Computer files (images, sounds recordings, even disks) contain unused or insignificant areas of data. Steganography takes advantage of these areas, replacing them with information (encrypted mail, for instance). The files can then be sent or transported without anyone knowing what really lies inside of them. An image of the space shuttle landing might contain a private letter to a friend. A recording of a short sentence might contain your company's plans for a secret new product. Steganography can also be used to place a hidden "trademark" in images, music, and software, a technique referred to as watermarking.
>
> ©1997, 1998 Eric Milbrandt

Steganography is not a commonly used word, but steganography is more commonly applied than is realized. These paragraphs are from a message on the cypherpunks mailing list:

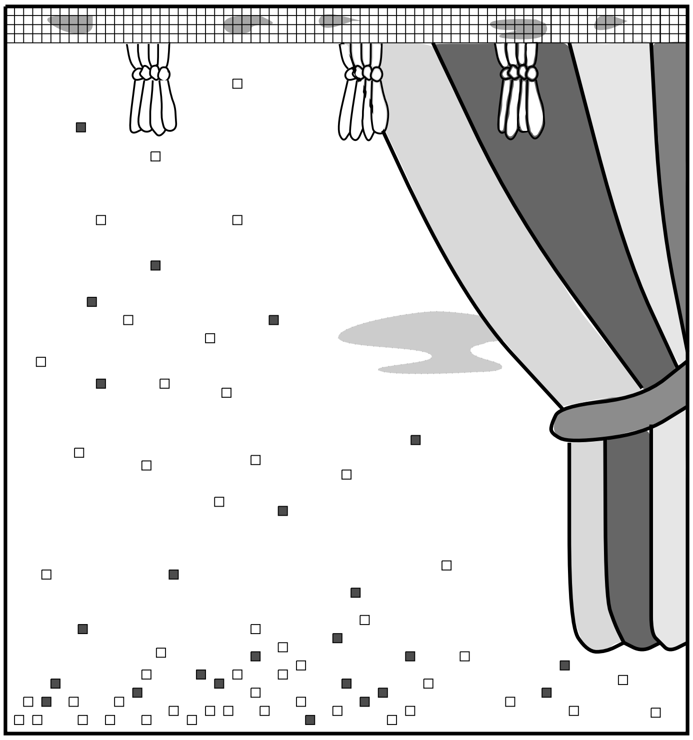
> ... when the potential for counterfeiting of valuable documents on color copiers/xerographic printers became apparent in Japan (where such machines first appeared) manufacturers were concerned about negative governmental reaction to such technology. In an effort to stave off legislative efforts to restrict such devices, various ID systems began being implemented at that point. At one stage for at least one U.S. manufacturer, this was as crude as a serial number etched on the underside of the imaging area glass!
>
> Modern systems, which are now reportedly implemented universally, use much more sophisticated methods, encoding the ID effectively as "noise" repeatedly throughout the image, making it impossible to circumvent the system through copying or printing over a small portion of the image area, or by cutting off portions of printed documents. ...
>
> To read these IDs, the document in question is scanned and the "noise" decoded via a secret and proprietary algorithm. In the case of Xerox-manufactured equipment, only Xerox has the means to do this, and they require a court order to do so (except for some specific government agencies, for whom they no longer require court authorizations). I'm told that the number of requests Xerox receives for this service is on the order of a couple a week from within the U.S.
>
> In the copier case, that ID technology being used for color copies *could* be adapted to black and white copies and prints as well. The addition of cheap GPS [global positioning

> systems: my note] units to copiers could provide not only valid date/time stamps, but also the physical *locations* of the units, all of which could be invisibly encoded within the printed images.
>
> Pressures to extend the surveillance of commercial copyright enforcement take such concepts out of the realm of science-fiction, and into the range of actual future possibilities.

There are many bits in most digital property whose alteration would hardly be noticed by a human observer. Huge music files have lots of room: the ear can hardly hear that well. And pictures ... so much can be hidden in even the simplest pictures.

A problem in steganography

Describe carefully how to hide 20 bits in the picture below, without harming its artistic *integrity or quality**. 20 bits is enough to encode more than a million distinct numbers since 2^{20} is 1,048,576.



BITSTORM

* Assuming there is any.

Lecture 7: More encryption

7.1 Diffie-Hellman key exchange

I'll briefly discuss the first published public key encryption system. It is simple to explain.

The goal is to have Alice and Bob share information that Eve does not know. We assume that Eve can "overhear" any exchange between them. The protocol described here gives a method for doing this which seems to be computationally secure, even when Eve has considerably superior computational resources.

Step 1: Alice and Bob prepare Alice and Bob together select a prime number P and another number, G, between 1 and P. Both P and G are public. Alice chooses a random private number a and Bob chooses a random private number b. a and b are *not* exchanged or made public.

Step 2: Alice and Bob privately compute Alice computes $A = G^a \bmod P$ and Bob computes $B = G^b \bmod P$.

Step 3: Alice and Bob exchange Alice sends A to Bob and Bob sends B to Alice. Now A and B are public.

Step 4: Alice and Bob again privately compute Alice computes $B^a \bmod P$ and Bob computes $A^b \bmod P$. Notice that $B^a = \left(G^b\right)^a = G^{ba}$ and $A^b = (G^a)^b = G^{ab}$, so $B^a = A^b$ (call this common value, K). Alice and Bob now share a common secret.

This all seems very strange. What does Eve know? She knows P and G and A and B: lots of information. She wants to know K.

7.2 An example

Here's a toy example so we can learn what Alice and Bob and Eve know.

Step 1: Alice and Bob prepare Alice and Bob choose the prime $P = 17$. They choose $G = 3$. Eve knows both of these numbers *and* knows how they will be used (Kerckhoff's maxim!). Secretly Alice chooses $a = 4$ and Bob chooses $b = 7$.

Step 2: Alice and Bob privately compute Alice computes $A = 3^4 = 13 \bmod 17$ and Bob computes $B = 3^7 = 11 \bmod 17$.

Step 3: Alice and Bob exchange Alice sends Bob the number 13, and Bob sends Alice the number 11. Eve now knows both 13 and 11.

Step 4: Alice and Bob again privately compute Alice computes $11^4 = 4 \bmod 17$ and Bob computes $13^7 = 4 \bmod 17$. Alice and Bob now have common information. Here $K = 4$.

Since Eve knows so much, she *must* know how to find K. In this case, she knows that $3^a = 14 \bmod 17$ and $3^b = 11 \bmod 17$. She needs to know $K = 3^{ab} \bmod 17$. (I will explain shortly how K is used cryptologically.) So the important implication to be investigated is:

If 3^a and $3^b \bmod 17$ are known, then what is the value of $3^{ab} \bmod 17$?

7.3 How hard is breaking Diffie-Hellman?

Since 3^a and 3^b are known, just take logs and divide by $\log 3$. That will give a and b. Then compute ab, exponentiate, and get K. Certainly that seems the naive answer. But remember that arithmetic is being done $\bmod P$ (in the toy example above, mod 13). With modular arithmetic, logs may not be as well-behaved as we'd like. The problem

Given a prime P and an integer G, describe
how to solve the equation $G^x = y \bmod P$.

when y is known is called the **discrete logarithm** problem. The assumption that this problem is "computationally infeasible" (in the language of an RSA, Inc. webpage!) underlies most of the belief in the security of the Diffie-Hellman protocol. This infeasibility has been tested by lots of smart people. As we have mentioned, exponentiating can be done "quickly" (in polynomial time). The inverse problem does not seem to be easy.

Also, using Diffie-Hellman is more complex than described. For example, a bad choice of G should be avoided. In the toy example, I chose G to be 3. The powers of 3 mod 17 are all of the distinct non-zero integers mod 17. In fact, the values of 3^n mod 17 are

$$3, 9, 10, 13, 5, 15, 11, 16, 14, 8, 7, 4, 12, 2, 6, 1$$

These are *all* of the non-zero numbers mod 17. If I had chosen $G = 4$, then the powers of 4 would just have given me 1 or 4 or 13 or 16: a smaller number of possible K's. G's should be chosen which have many different powers mod P. This can be done but is an additional complication.

7.4 Attacking the use of Diffie-Hellman: the man-in-the-middle

Diffie-Hellman and allied protocols are actually used in e-commerce. Here's a possible problem. Suppose Eve intercepts the communication between Alice and Bob. Eve can pretend to be Bob to Alice, and pretend to be Alice to Bob. Here's how.

We'll suppose G and P are known to everyone. We will also suppose that the connection between Alice and Bob is controlled by Eve, so Eve not only knows the information transmitted, but can modify or cancel it. Alice begins by sending $A = G^a \bmod P \longrightarrow ???$ Alice believes this message is going to Bob. Actually, Eve receives the message and prevents it from traveling further.

Eve invents her own e, computes $G^e \bmod P$ and sends this to Alice. Eve then takes Alice's A, computes $A^e \bmod P$ etc. Thus Eve (and, unwittingly, Alice!) share a *fake* key, K_{AE} which here is $G^{ae} \bmod P$. Eve can then pretend to be Bob to Alice. Eve can simultaneously pretend to be Alice to Bob with another K_{BE} obtained in a similar fashion. This process is easy when you know how and when all communication is done by remote digital signalling.

Eve's actions are called a man-in-in-the-middle attack. When you install a web browser, the web addresses of certain "trusted key authorities" are part of the browser's initial information to deter such attacks. These authorities are essentially lists of authentic public keys. The browser then can be led from one trusted authority to another (a sort of chain or pyramid of trust authorities), allowing its users to engage in secure internet commerce. Managing all this becomes quite intricate.

7.5 Bibliography

The original paper on Diffie-Hellman:

[1] W. Diffie and M. E. Hellman, *New directions in cryptography*, IEEE Transactions on Information Theory 22 (1976), 644-654. I have not been able to find a web version of this important paper.

[2] `http://cr.yp.to/patents/us/4200770/text` The text of the patent on the Diffie-Hellman cryptosystem: definitely reading for the overly dedicated!

[3] `http://www.alpertron.com.ar/DILOG.HTM` This is a webpage with a Java applet to compute discrete logarithms.

[4] John Franco, a webpage at the University of Cincinnati explaining Diffie-Hellman:
`http://gauss.ececs.uc.edu/Users/Franco/Project/dh.htm`
with another page allowing small examples to be tried:
`http://gauss.ececs.uc.edu/Users/Franco/Project/dh2.htm`

[5] There are RSA challenges, usually having to do with factoring. Similarly, there are discrete logarithm challenges (with large monetary prizes). See, for example, the web page at Certicom: `http://www.certicom.com/resources/ecc_chall/challenge.html`. The Certicom web page remarks (about both factoring and discrete log):

> None of these problems have been proven to be intractable (i.e., difficult to solve in an efficient manner). Rather, they are believed to be intractable because years of intensive study by leading mathematicians and computer scientists around the world has failed to yield efficient algorithms for solving them. As more effort is expended over time in studying and understanding these problems, our confidence in the security of the corresponding cryptographic systems will continue to grow.

Does this encourage confidence?

Lecture 8: Perfect cryptography

8.1 Mod 2

First, a small digression on addition mod 2. This is sometimes called **xor** (for exclusive or) and written with the symbol $\oplus$. The entire definition of $\oplus$ is given by these equations:

$$0 \oplus 0 = 0; \quad 0 \oplus 1 = 1; \quad 1 \oplus 0 = 1; \quad 1 \oplus 1 = 0.$$

If 0 represents True and 1, False, you can probably see the reason for the phrase "exclusive or". $\oplus$ is used extensively in cryptography because $a \oplus b \oplus b = a$ for any a's and b's, so we can do the following:

$$\overbrace{a}^{\substack{\text{One original}\\ \text{message bit}}} \longrightarrow \overbrace{a \oplus b = m}^{\substack{\text{Encryption}\\ \text{using } b}} \longrightarrow \overbrace{m \Rightarrow m}^{\substack{\text{Message}\\ \text{transmission}}} \longrightarrow \overbrace{m \oplus b = a \oplus b \oplus b = a}^{\substack{\text{Decryption}\\ \text{using } b}} \longrightarrow \overbrace{a}^{\substack{\text{Original message}\\ \text{bit received.}}}$$

Here's a *provably* perfect encryption method. Convert your message into a stream of bits: just 0's and 1's. Take a random bitstream, and xor your message bitwise (the first message bit xor'd with the first random bit, the second message bit xor'd with the second random bit, etc.). Then send the stream of resulting bits. Your receiver can decrypt by xoring with the same random bitstream.

What's "random"? By random I mean here that I don't know and can't predict the bits, that you don't know and can't predict the bits, and that no one knows or can predict the bits in the random bitstream. That's my "definition". Now I'll prove that the encryption method is perfect. If you could get the message bit, a, from the encrypted bit, m, then since $m = a \oplus b$ and you know m and a, the equation $a \oplus m = a \oplus a \oplus b = b$ shows that you would know b. But we remarked that you did not know the b's. This is a contradiction, so you can't get the message bit from the encrypted bit.

8.2 Comments, questions, ...

What's described resembles the book code discussed in the section whose title is "Classical crypto". Here the alphabet isn't English, but just 0 and 1. The commonly used name for this encryption method is the **one-time pad**. If the same stream of random bits is used twice, then Eve would know $m_1 = a_1 \oplus b$ and $m_2 = a_2 \oplus b$. But then she would also know $m_1 \oplus m_2 = a_1 \oplus a_2$, since the b's xor'd twice cancel. But bits of original messages are not random because languages have a great deal of statistical irregularity. Therefore it is frequently quite possible to unravel two original messages only given the xor of these messages.

How can random bitstreams be created? Natural phenomena can be used (radioactivity, heat, etc.). This can be very clumsy and actually quite difficult (but see [1] for your very own random bits, please!). Then the random bitstreams need to be packaged somehow, and duplicates made for the sender and the receiver. Such "key management" can be quite complicated and very difficult. Every possible pair of senders and receivers would need distinct random bitstreams. Maybe now you can begin to understand the complex history of secure diplomatic, military, and commercial communication. Also, lots and lots and lots of bits are needed to send music and pictures and movies etc. And people want all of this done very rapidly. Even with fast computation, public key exchanges take a while.

8.3 What is actually done

"Keys" are exchanged using an initial public key negotiation. These keys then are used to initialize pseudo-random number generators. I'll describe the generator used by `Maple`. Although `Maple` has numerous ways of producing "random" objects, all of them depend on the following idea.

- Start with some integer x_0. Unless told otherwise, `Maple` takes x_0 to be 1.
- Produce a sequence of "random" integers $\{x_n\}$ by using the transformation $x_{n+1} = cx_n \bmod m$ where $c = 42\,74196\,69081$ and $m = 99\,99999\,99989$.

This method is called a linear congruential generator. c and m were very carefully selected so that the cycle length (the number of x_n's before there is repetition) is actually $m - 1$, a rather large number. The initial integer, x_0, can be chosen by the user with the command `_seed:=`the user's value. Although `Maple`'s choice of a pseudo-random number generator is fine for many purposes, cryptography has stricter requirements. Linear congruential generators can be detected rather easily.

Most web browsers have implementations of an algorithm called RC4. Secure web traffic is initiated by a public key exchange using one of the protocols mentioned above (usually mediated by a trusted authority, to prevent man-in-the-middle attacks!). Then the key that was exchanged is used to start RC4, which is much more complicated than a linear congruential generator. One citation at RSA, Inc., declares, "Analysis shows that the period of the cipher [RC4] is overwhelmingly likely to be greater than 10^{100}*. Eight to sixteen machine operations are required per output byte, and the cipher can be expected to run very quickly in software."

Much research into "randomness" has been done (and is in progress!). Probably we should investigate what random and non-random really mean.

8.4 Bibliography

[1] "True" random numbers are available at `http://www.random.org` which was created by Mads Haahr. The essay `http://www.random.org/essay.html` has much information, including further links. You can also look at `http://www.random.org/users.html` to see some of the uses of these really random numbers.

* That number is indeed big enough for most interactions. [My comment!]