How to Break MD5 and Other Hash Functions

MD5是论文发表时应用最广泛的密码哈希函数之一，在本文中，王小云教授团队提出了一个新的差分攻击算法，能快速有效地找到MD5的碰撞。

MD5

在此论文中对MD5的介绍并非十分详尽，我参考在网上找到的其他资料作为补充，对MD5算法进行了了解和学习。

MD5简介

MD5的全称是Message-Digest Algorithm，翻译过来叫消息摘要算法，由美国密码学家罗纳德·李维斯特（Ronald Linn Rivest）设计。 消息摘要算法是密码学算法中非常重要的一个分支，它通过对所有数据提取指纹信息以实现数据签名、数据完整性校验等功能，由于其不可逆性，有时候会被用做敏感信息的加密。摘要算法的目标是用于证明原文的完整性，也就是说用于防止信息被篡改。消息摘要算法也被称为哈希（Hash）算法或散列算法。它是一种密码散列函数，任何消息经过散列函数处理后，都会获得唯一的散列值，这一过程称为 “消息摘要”，其散列值称为 “数字指纹”，如果数字指纹一致，就说明消息是一致的。
使用MD5算法进行消息的加密和验证，主要分为这几个步骤：

1. 发送方使用MD5算法为一段消息生成独一无二的MD5摘要，作为这段消息的签名。
2. 发送方将欲发送的消息和MD5摘要一起发送给接收方。
3. 接收方收到消息后，使用MD5算法生成收到的消息的MD5摘要，然后将生成的这个MD5摘要和接收到的MD5摘要进行对比，如果两份摘要一致，则认为接收到的消息完整，没有被篡改，如果不一致，则认为消息受到篡改，要求发送方重新发送消息。

MD5算法生成MD5摘要

MD5以512位分组来处理输入的信息，且每一分组又被划分为16个32位子分组，经过了一系列的处理后，算法输出一个由四个32位分组组成的128位散列值。

1. 填充：MD5算法是对输入的数据进行补位，使得如果数据位 长度LEN对512求余的结果是448。即数据扩展至K}512+448 位。即K*64+56个字节，K为整数。具体补位操作：补一个 1，然后补0至满足上述要求。
2. 记录信息长度：用一个64位的数字表示数据的原始长度B，把B用两个 32位数表示。这时，数据就被填补成长度为512位的倍数。
3. 初始化MD5参数：四个32位整数(A，B，C，D)用来计算信息摘要，初始化使用的是十六进制表示的数字：A=0X0123456, B=0X89abcdef, C=0Xfedcba98, D=0X76543210
4. 处理位操作函数：X, Y, Z为32位整数。 F(X, Y, Z) = X&Y|NOT(X)&Z; G(X，Y, Z) = X&Z|Y&NOT(Z); H(X, Y, Z) = X xor Y xor Z; I(X, Y, Z) = Y xor (X|NOT(Z))
5. 四轮变换：使用常数组T[1...64]，T为32位整数用16进制表示，数据用16个32位的整数数组M[]表示。对原文进行处理之后存放在16个元素的数组X中。定义[abcd k s i]表示操作a = b + ((a + F(a, b, c) + X[k] + T) <<< s)，对ABCD进行四轮每轮包含16次此操作的变换，最后得到MD5摘要结果。

碰撞

在本文中，王小云教授描述了一个可以轻易构造MD5碰撞实例的算法。 在计算机科学中，碰撞是指两个不同的元素具有相同的哈希值、校验和，数字指纹时发生的情况，即指不同的输入却产生了相同的输出，当数据量足够多（例如将所有可能的人名和计算机文件名映射到一段字符上）时，碰撞是不可避免的。 比如下面两组不完全相同的数据：

```
d131dd02c5e6eec4693d9a0698aff95c
2fcab58712467eab4004583eb8fb7f89
55ad340609f4b30283e488832571415a
085125e8f7cdc99fd91dbdf280373c5b
d8823e3156348f5bae6dacd436c919c6
dd53e2b487da03fd02396306d248cda0
e99f33420f577ee8ce54b67080a80d1e
c69821bcb6a8839396f9652b6ff72a70
```

和

```
d131dd02c5e6eec4693d9a0698aff95c
2fcab50712467eab4004583eb8fb7f89
55ad340609f4b30283e4888325f1415a
085125e8f7cdc99fd91dbd7280373c5b
d8823e3156348f5bae6dacd436c919c6
dd53e23487da03fd02396306d248cda0
e99f33420f577ee8ce54b67080280d1e
c69821bcb6a8839396f965ab6ff72a70
```

有着相同的MD5摘要：`79054025255fb1a26e4bc422aef54eb4`

差分攻击

差分攻击是对MD5这样的哈希函数的最重要的分析方法，分析分组密码的一种重要方法。差分密码分析通常是选择明文攻击，意思是攻击者可以自行选取一部分明文并获取相应密文。不过，一些扩展也能让此方法用在已知明文攻击，甚至是唯密文攻击上。差分分析的基本方法，是运用若干对有着常量差异的明文；差异可以用数种方法定义，最常见的是逻辑异或（XOR）。然后，攻击者计算相应密文的差异，尝试找出差异分布的统计特征。明文差异和密文差异所组成的差异对被称为差分，其统计性质由加密所使用的S盒决定。也就是说，对于S盒子S，攻击者分析差分(ΔX, ΔY)，其中ΔY = S(X ⊕ ΔX) ⊕ S(X)（⊕表示异或）。在初等攻击中，攻击者希望某个密文差异出现的频率非常高，这样就能将加密和随机过程区分开来。更复杂的变体攻击能做到比穷举更快地破解出密钥。

MD5差分攻击

本文中，王小云教授使用模整数方法的差分攻击，利用压缩模块中的缺陷，证明了标准MD5算法的碰撞不稳固。本文提出了充分条件的概念，并列出了一系列的充分条件（大约有290个），如果这些充分条件都能得到满足，那么一定能产生碰撞。于是MD5的强抗碰撞性不能得到满足，即该攻击方法可以寻找消息对(x, y)，使得MD5(y) = MD5(x)。不过，这一系列的充分条件很难同时满足。尽管本文等进一步提出了消息修改算法，通过修改相应比特位的方法来达到满足这一系列充分条件，但是仍然有37条充分条件不能满足。这就意味着，从理论上来讲，该算法只需测试2^37条随机消息就可以找到完全满足充分条件的消息对(x, y)，从而找到

碰撞，即MD5(y) = MD5(x)。
本文在p690系列机上进行实验，只需15分钟到1小时的计算时间就能寻找到MD5碰撞，而相比于p690的1,025,486tpmC，目前最领先的机器阿里巴巴的OCEAN的性能是他的700多倍，有707,351,007tpmC，也就是说目前的MD5碰撞能在更短的时间内完成。

不足

王小云教授的算法提供了迅速找到了MD5碰撞的方法，证明了MD5算法的不安全性，对数学和密码学做出了很大的贡献，但是也有它的局限性。首先，此方法找到的的碰撞是随机的，其次，找到碰撞还远远不能称之为破解，只是找到了与原始信息MD5摘要相同的信息，离实用价值还有一些距离。