Detecting Spam with Genetic Regular Expressions

GCIA Gold Certification

Author: Eric Conrad, eric@ericconrad.com

Adviser: Johannes Ullrich

Accepted:

Eric Conrad
Detecting Spam with Genetic Regular Expressions

Outline

1. Abstract ............................................................................................................................... 3
2. Introduction .......................................................................................................................... 4
3. An Approach to Automatically Detect Spam with Genetic Regular Expressions .............. 11
4. Genetic Regular Expressions ............................................................................................... 18
5. Proof-of-Concept Script Results ......................................................................................... 29
6. Conclusion .......................................................................................................................... 43
7. Further study ....................................................................................................................... 44
8. Appendices ......................................................................................................................... 46
9. References ......................................................................................................................... 66
1. Abstract

Regular Expressions ("regex" for short) are strings used to detect patterns in data. They are often used to detect and block various forms of malware, including spam and network-based attacks.

This paper describes an approach for detecting spam with automatically-generated regular expressions (where regexes are generated according to simple logic), followed by a 'genetic' approach (where regexes are generated, and then 'evolve' to the final solution via a genetic algorithm).

This approach was inspired by the author's experience running inbound internet SMTP gateways for a 12,000 employee company. As of the fall of 2007, the gateways blocked 150,000 spam per business day.
2. Introduction

Problem Domain

The ‘problem domain’ (specific challenge) addressed by this paper is detecting spam emails via genetic regular expressions, focusing on ‘Advanced Fee’ scams. These scams include lottery scams (where the target is told he/she has won a lottery, and is asked to submit a fee to receive payment), and ‘419’ scams.

A 419 scam is a scam where an alleged foreign government official requests an upfront payment in exchange for releasing large sums of money. It is named after the section of Nigerian legal code violated by the scam\(^1\).

The FBI’s article titled ‘Common Fraud Schemes’ describes the scam:

\[\begin{align*}
\text{Payment of taxes, bribes to government officials, and legal fees are often} \\
\text{described in great detail with the promise that all expenses will be reimbursed as soon} \\
\text{as the funds are spirited out of Nigeria. In actuality, the millions of dollars do not exist} \\
\end{align*}\]

\(^1\) According the International Centre for Nigerian Law: ‘Any person who by any false pretence, and with intent to defraud, obtains from any other person anything capable of being stolen, or induces any other person to deliver to any person anything capable of being stolen, is guilty of a felony.’ URL: http://www.nigeria-law.org/Criminal%20Code%20Act-Part%20VI%20%20to%20the%20end.htm

Eric Conrad
Detecting Spam with Genetic Regular Expressions

and the victim eventually ends up with nothing but loss. Once the victim stops sending money, the perpetrators have been known to use the personal information and checks that they received to impersonate the victim, draining bank accounts and credit card balances until the victim's assets are taken in their entirety. While such an invitation impresses most law-abiding citizens as a laughable hoax, millions of dollars in losses are caused by these schemes annually. Some victims have been lured to Nigeria, where they have been imprisoned against their will, in addition to losing large sums of money. The Nigerian government is not sympathetic to victims of these schemes, since the victim actually conspires to remove funds from Nigeria in a manner that is contrary to Nigerian law. The schemes themselves violate section 419 of the Nigerian criminal code, hence the label "419 fraud."\(^2\)

The scam originated in Nigeria, but is now used around the world, describing various government and private sector officials with alleged access to large sums of money.

In the fall of 2007, the most spam complaints received at my company of 12,000 employees involved 419 scams. The inbound internet mail relays run Postfix, SpamAssassin, and antivirus software, and do an excellent overall job of blocking most types of spam.

\(^2\) URL: [http://www.fbi.gov/majcases/fraud/fraudschemes.htm](http://www.fbi.gov/majcases/fraud/fraudschemes.htm)

Eric Conrad 5
Detecting Spam with Genetic Regular Expressions

Here's an example 419 scam email received 'in the wild':

To: 
Subject: Hello
From: "villaran nenita" <villaran1976_n@citromail.hu>
Date: Thu, 13 Sep 2007 12:16:56 CEST

Hello Dear,

My name is nenita villaran, The wife of Mr. Panfilo Nenita a senator during president joseph Estrada regime in Philippine. who is recently killed in philippine. During my husband's regime as a senator, I realized some reasonable amount off money from various deals that I successfully executed and my buisness in the united state of which the Government has block most of my account in the bank of america, trying to leave me with nothing.

well before my late husband was killed, I secretly put in a box the sum of $30,000,000 million USD (Thirty million United states dollars) and deposit it in a security company abroad. I am contacting you because I want you to help me in securing the money for the future of my children since the government now monitor all my movement.

I hope to trust you as who will not sit on this money when you claim it. I will give you 15% of the total money for your assistance. If you are willing to help me as soon as possible.

Best regards=20

Villaran Nenita

The challenge in filtering these types of emails with software such as SpamAssassin is that they are comprised mainly of English text, designed to look somewhat like a business letter. They may be sent from any domain (no forging required), and do not typically contain a trackable URI, which may be used to identify spam via 'URI Blacklists.'

This specific spam was detected by these SpamAssassin rules (among others):

___FRAUD_JYG ======> got hit: "give you 15% of the total"

---

3 Personal email communication, September 17, 2007
4 One example is the URI BL. URL: http://www.uribl.com/about.shtml

Eric Conrad
Detecting Spam with Genetic Regular Expressions

Here’s a 419 scam email that was not blocked by SpamAssassin:

From: Fund Manager  
Sent: Friday, September 07, 2007 9:46 AM  
To: undisclosed-recipients  
Subject: Mail..  
Re: Confidential Deal:

Greetings,

I am a Fund Manager with Fidelity Investment UK and I handle all our Investor’s Capital Project Funds that enables me to divert 1.2% Investors Excess Return Capital Funds to our Magellan Trust Funds Account whereby anyone can be presented to claim the funds. On this note, the total sum of US$34.5M has been diverted representing the 1.2% Excess Return Capital Funds from the Investor Capital Project Funds for 2006/2007. I need a reliable and trustworthy person that can work this deal out with me so that we can claim the funds as mentioned above.

There is no risk attached and the funds in question can never be dictated or traced. Our sharing ratio is 50:50. If you are interested, please send your direct telephone numbers for discussion of this deal in further details. I am counting on your sense of confidentiality, as it is my desire that you keep this business to yourself.

Sincerely.

Mr. Carlos Moreno.  

It triggered these SpamAssassin rules:

-FRAUD_YPO ======> got hit: "the total sum"  
-FRAUD_IOU ======> got hit: "no risk"

Although it triggered some rules, the overall SpamAssassin score was under the site’s quarantine threshold, so the email was passed onto the end user (who later complained to the internet mail team).

5 Output generated with ‘spamassassin –D’, showing lines with ‘FRAUD’. Cleaned up for formatting purposes
6 Personal email communication, September 3, 2007

Eric Conrad
Detecting Spam with Genetic Regular Expressions

Regular Expressions

Regular expressions are a powerful mechanism for matching text. A simple regular expression matches literal text. Regxes may also contain advanced features such as metacharacters, character classes, and grouping (among others).

Metacharacters are concise commands which perform a function, such as ‘match the start of line’ (“^”). Character classes allow a range of characters, such as [a-z] for lowercase characters. Grouping allows a choice of words, like “(FEE|FIE|FOE|FUM)”.

This example uses all three types:

/^You are in a maze of twist(y|ing) little passages[, .]/

This regex matches lines beginning with “You are in a maze of ...”, allows the choice of ‘twisty’ or ‘twisting’, and matches a space, period, or comma after ‘passages’.

Perl Compatible Regular Expressions (PCRE) are a powerful regular expression engine derived from the Perl programming language. Many software packages use PCREs,

_____________________

7 /This is a literal match/
9 PCRE - Perl Compatible Regular Expressions. URL: http://www.pcre.org/
10 Information about Perl may be found at The Perl Directory at Perl.org. URL: http://www.perl.org

Eric Conrad 8
Detecting Spam with Genetic Regular Expressions

including SpamAssassin, Snort, Apache, Postfix, and many others. This paper uses PCREs in all examples and code.

Leveraging Regular Expressions to Block Malware

Regular expressions are harnessed in many types of software, including intrusion detection systems and spam-blocking software.

Here’s an example of a Snort rule including a PCRE:\textsuperscript{11}

\begin{verbatim}
alert tcp $EXTERNAL_NET 22 -> $HOME_NET any (msg:"EXPLOIT SSH server banner overflow"; flow:established,from_server; content:"SSH-"; nocase; isdataat:200,relative; pcre:"/^SSH-\s[\n]{200}/ism"; reference:bugtraq,5287; reference:cve,2002-1059; classtype:misc-attack; sid:1838; rev:8;)
\end{verbatim}

A simplified description of this rule is that it will match on traffic with a TCP source port of 22 (the SSH daemon port), matching the PCRE "/^SSH-\s[\n]{200}/ism".

That means: match the literal string “SSH-“ at the beginning of the line, followed by at least 200 characters that are not newlines. The search is case insensitive (‘i’), will match across multiple lines (“m”), and ‘.’ (‘dot’ metacharacter which matches any character) will also match a newline (“s”).

\textsuperscript{11} Official Snort 2.4 rules. URL: http://www.snort.org/pub-bin/downloads.cgi/Download/vrt_pr/snortrules-pr-2.4.tar.gz, /rules/exploit.rules

Eric Conrad
Detecting Spam with Genetic Regular Expressions

Here is an example of a SpamAssassin rule containing a regex:

```
Subject =~ /(Your Lucky Day|Attention:|ONLINE) WINNER)/i
```

That matches a Subject which includes any of the following:

- Your Lucky Day
- Attention: WINNER
- ONLINE WINNER
- OnLiNe wInNeR

That last example matches because the search is case sensitive.

---

12 SpamAssassin 3.2.3. URL: http://mirror.fslutd.org/apache/spamassassin/source/Mail-SpamAssassin-3.2.3.tar.gz, Mail-SpamAssassin-3.2.3/rules/72_active.cf
3. An Approach to Automatically Detect Spam with Genetic Regular Expressions

Bayesian Filtering with Spam and Ham

While regular expressions are typically written by humans, there are classes of software that use statistical methods to automatically detect spam. One such approach is Bayesian filtering, named after Thomas Bayes, an English clergyman who devised a number of probability and statistical methods, including “a simple mathematical formula used for calculating conditional probabilities.”

Paul Gram described Bayesian filtering to identify spam in his paper ‘A Plan for Spam.’ He described using a ‘corpus’ of ‘spam’ and ‘ham’, human-selected groups of spam and non-spam., respectively. He then used Bayesian filtering techniques to automatically assign a mathematical probability that certain ‘tokens’ (words in the email) were indications of spam.

We will use the spam and ham corpus approach to identify the fitness of genetic

14 URL: http://www.paulgraham.com/spam.html
Detecting Spam with Genetic Regular Expressions

regular expressions.

Automatic Regular Expressions

Regular expressions leveraged to block spam or malware may be automatically generated according to the following logic:

1. Create 2 sets of emails called 'spam' and 'ham'
2. Break spam into tokens
3. Generate a large number of regular expressions based on the spam tokens
4. Delete regex that match any ham
5. Delete regex that match only 1 spam
6. Rank the remaining regex, sorted from highest number of matching spam

The regular expressions are generated using the following algorithm, applied in order (first match wins):

- If its a number, replace with \d+ (1 or more digits)
- If its capital hexadecimal, replace with [A-F0-9]+ (1 or more capital hex digits)
- If its lowercase hexadecimal, replace with [a-f0-9]+ (1 or more lowercase hex digits)
- If its a common TLD, replace with (com|net|org|edu|biz|info|us)
- If its a lowercase word, replace with [a-z]+ (one or more lowercase letters)
- If its an uppercase word, replace with [A-Z]+ (one or more uppercase letters)
- If its an abbreviated day of the week, replace with (Mon|Tue|Wed|Thu|Fri|Sat|Sun)
Detecting Spam with Genetic Regular Expressions

- If its an abbreviated month, replace with (Jan|Feb|Mar|Apr|May|Jun|Jul|Aug|Sep|Oct|Nov|Dec)
- If its a word with the first letter capitalized, replace with [A-Z][a-z]+ (uppercase letter, followed by one or more lowercase letters).
- Else: keep the token as a literal regex.

Please see Appendix A for proof-of-concept Perl source code of the above logic.

These rules were run against a small corpus of spam (339) and ham (410) from personal email. Some spam were previously tagged with "***SPAM***" in the subject line (by SpamAssassin); the tag was left in to determine if any regexes would emerge that matched that tag.

The following regular expressions were generated, ranked in order of spam matched:

```
<table>
<thead>
<tr>
<th>Count</th>
<th>Regex</th>
</tr>
</thead>
<tbody>
<tr>
<td>22</td>
<td>/^From: &quot;[A-Z][a-z]+ [A-Z][a-z]+&quot; &lt;[a-z]+@[a-z]+.[a-z]+.[a-z]+&gt;/</td>
</tr>
<tr>
<td>27</td>
<td>/^Subject: [a-z]+ [a-z]+ [a-z]+&gt;/</td>
</tr>
<tr>
<td>32</td>
<td>/^Subject: [a-z]+ [a-z]+&gt;/</td>
</tr>
<tr>
<td>33</td>
<td>/  +[a-z]+=&quot;----=<em>NextPart</em>\d+_[A-F0-9]{8}.[A-F0-9]{8}&quot;/</td>
</tr>
<tr>
<td>37</td>
<td>/^Subject: &quot;***[A-Z]+***[A-Z]+ [a-z]+ [a-z]+ [a-z]+&gt;/</td>
</tr>
<tr>
<td>43</td>
<td>/^Subject: &quot;***[A-Z]+***[A-Z]+ [a-z]+ [a-z]+&gt;/</td>
</tr>
<tr>
<td>52</td>
<td>/  +[a-z]+=&quot;-----<em>NextPart</em>\d+<em>[A-F0-9]{4}</em>[A-F0-9]{8}&quot;/</td>
</tr>
<tr>
<td>57</td>
<td>/^Subject: &quot;***[A-Z]+***[A-Z]+ [a-z]+&gt;/</td>
</tr>
<tr>
<td>61</td>
<td>/^From: &quot;[A-Z][a-z]+ [A-Z][a-z]+&quot; &lt;[a-z]+@[a-z]+(com</td>
</tr>
<tr>
<td>88</td>
<td>/^From: &quot;[A-Z][a-z]+ [A-Z][a-z]+&quot; &lt;[a-z]+@[a-z]+.[a-z]+&gt;/</td>
</tr>
</tbody>
</table>
```

"\*\*\*[A-Z]+\*\*\*[A-Z]+" is a regex that matches "***SPAM***": three asterisks ("*"); followed
Detecting Spam with Genetic Regular Expressions

by an uppercase word (1 or more of the capital letters A-Z), followed by three asterisks.\textsuperscript{15}

Here are some matches:

Subject: ***SPAM*** And who pronunciation
Subject: ***SPAM*** Are lacamp he estherville
Subject: ***SPAM*** Are well coahoma
Subject: ***SPAM*** As she kennedyville

This tag (inserted by SpamAssassin) was left in the spam corpus as a test, to see if
regex would be generated which match it.

The regex:

\[
/\ +[a-z]+="-----_NextPart_\d+_\{A-F0-9\}{4}_{A-F0-9}\{8\}.\{A-F0-9\}\{8\}"/
\]

...matches a MIME (Multipurpose Internet Mail Extension) attachment headers:

boundary="-----_NextPart_000_000F_01C78661.D9DA98C0"
boundary="-----_NextPart_000_000F_01C78818.3AAE1BB0"
boundary="-----_NextPart_000_000F_01C788D2.D850C470"
boundary="-----_NextPart_000_000F_01C78EAE.D04A6A60"

They appear to be legitimate MIME boundaries to the naked eye. The regex ‘ +’
means ‘match two or more spaces.’ The ham corpus contained similar headers, but those
began with a leading ‘tab’ character (\texttt{\textbackslash t}). Some mail in the spam corpus replaced the tab with

\texttt{\textbackslash t}\texttt{\t}

\textsuperscript{15} Note that ‘*’ is a metacharacter (which means ‘match zero or more of the previous character’) which must be

Eric Conrad
Detecting Spam with Genetic Regular Expressions

8 leading spaces, which was a reliable indicator of spam within the corpus.

The automatic method of generating regexes appears to be effective.

Genetic Algorithms

A Genetic algorithm (GA) is an algorithm that is automatically generated, and then 'bred' through multiple generations to improve it via Darwinian principles:

*Genetic algorithms are search algorithms based on the mechanics of natural selection and natural genetics. They combine survival of the fittest among string structures with a structured yet randomized information exchange to form a search algorithm with some of the innovative flair of human search. In every generation, a new set of artificial creatures (strings) is created using bits and pieces of the fittest of the old; an occasional new part is tried for good measure. While randomized, genetic algorithms are no simple random walk. They efficiently exploit historical information to speculate on new search points with expected improved performance.*

‘Genetic regular expressions’ may be developed by combining genetic algorithms, the escaped with a backslash in order to be treated as a literal asterisk.


Eric Conrad
Detecting Spam with Genetic Regular Expressions

spam and ham corpus approach used in Bayesian filtering, and automatic regular expressions described in the previous section. The goal is to solve the complex problem of matching 419 spam text. As genetic algorithm pioneer John Holland said “Computer programs that ‘evolve’ in ways that resemble natural selection can solve complex problems even their creators do not fully understand.”

Key GA Terms

Genetic algorithms use chromosomes made up of individual genes. Chromosomes are designed to perform a specific function, such as sorting numbers, playing tic-tac-toe, or, in this case, matching spam text.

Chromosomes are scored according to a fitness function. Higher-scoring chromosomes are more likely to survive and are able to breed future generations. Breeding includes mating with other chromosomes and exchanging genes via a ‘crossover’ function. Stronger chromosomes are more likely to be chosen to breed. One way to select parents is ‘Roulette Wheel’ selection, described below.

Mutation may occur, where a random change is made to the chromosome. Many


Eric Conrad

© SANS Institute 2007, As part of the Information Security Reading Room Author retains full rights.
mutations may be damaging, but some could improve the 'health' of the chromosome.

A typical genetic algorithm uses the following logic:\(^\text{18}\)

- Create a large number of chromosomes
- Loop until complete:
  - Have the chromosomes perform a specific function
  - Apply a fitness score to each chromosome
  - Identify the strongest chromosomes and allow them to survive to the next generation
  - Breed new chromosomes by selecting genes from 2 parents
  - Mutate some genes

The loop ends at a fixed time (100 generations, for example), or when a certain condition is met, such as laying a perfect game of tic-tac-toe, or sorting integers in a specific number of steps.

\(^\text{18}\) For more background on practical genetic algorithms, including detailed pseudocode, see Dracopoulos, Dimitris C. 2A\text{IT608} - Machine Learning Genetic Algorithms. URL:
http://users.wmin.ac.uk/~dracopd/DOCUM/courses/2ait608/genetic_algorithms_lecture_notes.pdf
4. Genetic Regular Expressions

A regular expression chromosome is a full regular expression. Genes within the chromosome are literals, strings, metacharacters, and character classes (or groups thereof).

Here is an example regular expression chromosome:

```
Subject: /*\*\*[A-Z]+/\*\*\* [A-Z][a-z]+ [a-z]+/
```

Here are the individual genes:

1. Subject:
2. /*\*\*
3. [A-Z]+ 
4. /*\*\*
5. [A-Z][a-z]+ 
6. [a-z]+

The following process is used to evolve genetic regular expressions:

1. Generate a population of regex chromosomes based on the automatic method described above.
2. Match them against the spam and ham corpora\(^{19}\)
3. Assign a fitness score
4. Choose the fittest chromosomes to survive to the next generation (the rest ‘die’)
5. Crossover genes between some parents to produce children
6. Mutate genes in some children

\(^{19}\) The plural of ‘corpus’ is ‘corpora,’ although ‘corpuses’ may also be used. See: http://wiki.apache.org/spamassassin/PluralOfCorpus?highlight=%28ham%29%7C%28corpus%29

---

Eric Conrad
Detecting Spam with Genetic Regular Expressions

7. Repeat process for X generations

Genetic Regular Expressions Psuedocode

The full Perl source code to create genetic regular expressions is included in Appendix

B. Here is the same logic, in pseudocode:

1. Choose the number of generations (we’ll use 10)
2. Read the spam and ham corpora
3. Randomly shuffle the lines of spam
4. Divide spam corpus into 10 slices
5. Loop until 10th generation:
   a. Generate chromosomes based on the current slice of spam using the ‘automatic’ formula
   b. Score chromosomes
   c. Print results for the current generation
   d. Keep the fittest 3rd
   e. Breed survivors
      i. 2 survivors breed via a crossover function to create a child
      ii. Use ‘Roulette Wheel’ selection top choose the 2nd parent
   f. Mutate some of the children by randomly deleting some genes
   g. Move to next slice of spam
6. Print Final results

Sources of ham and spam

The ham used in this paper is based on the SpamAssassin ‘easy ham’ corpus,

Eric Conrad
Detecting Spam with Genetic Regular Expressions

downloaded from the SpamAssassin Public corpus. The spam corpus was collected from my company’s inbound internet relays from September 12th-14th, 2007.

The SpamAssassin public ham corpus was chosen due to its high quality, and because it was pre-cleaned of any sensitive information (such as names, IP addresses, etc). Current ‘advanced fee’ spam was chosen because it is required for addressing the problem domain of blocking ‘advanced fee’ scam emails currently bypassing spam filters. Also, the spam bodies contain no sensitive information.

The spam and ham corpora used in this paper are publicly available, along with the POC code. Please see the appendices for more information,

Both corpora were formatted according to the following logic:

1. Remove the headers
2. Remove all lines not consisting solely of the following characters:
   a. Alphanumeric
   b. Whitespace
   c. Comma
   d. Period
   e. Underscore
3. ‘sort’ and ‘unique’ the lines

The goal is to isolate lines containing human-written text, and skip headers, MIME

20 URL: http://spamassassin.apache.org/publiccorpus/20030228_easy_ham.tar.bz2

Eric Conrad
Detecting Spam with Genetic Regular Expressions

information, etc.

**Automatic chromosomes**

The proof-of-concept scripts seeds each round with chromosomes based on the spam corpus. The following logic is used in order to introduce randomness in the seeding:

For each prospective gene, randomly choose one of the following:

1. Leave it as a literal regex
2. Change it to a regular expression, using the ‘Automatic Regular Expression’ rules described previously
3. Change it to ‘.*’, meaning ‘match zero or more of any character’

Convert any whitespace to “\s+”\(^{22}\). Then combine any repeated “.*”\(^{23}\)

During the example run below, this line of spam:

```
me to Register it as Consignment with GLOBAL TRUST COURIER COMPANY.
```

...was converted to this chromosome, using the above logic:

```
/^[a-z]+\s+to\s+[A-Z][a-z]+\s+[a-z]+\s+Consignment\s+[a-z]+\s+.\*[A-Z]+\s+COU\nRIER\s+/\n```

\(^{21}\) On Unix, ‘sort –u <unsortedspam.txt >sortedspam.txt’

\(^{22}\) ‘match 1 or more whitespace characters’

\(^{23}\) ‘.*.’ is the same as ‘.*’, but less efficient.
Detecting Spam with Genetic Regular Expressions

Fitness Function

Chromosomes that match any ham ‘die’ (are ignored by the program). The surviving chromosomes are then scored by the fitness function.

The fitness function has 2 components: number of spam lines matched, and length of the chromosome. Short chromosomes that match lots of spam receive high scores; long chromosomes that match little spam receive low scores.

Remaining chromosomes are scored according to this fitness function:

\[
(Number\ of\ lines\ of\ spam\ matched) \times \left(1 + \frac{200 - \text{length of chromosome}}{200}\right);
\]

In other words, any chromosome under 200 characters long will receive a fitness bonus; chromosomes over 200 characters receive a penalty.

A 100 character chromosome matching 5 lines will receive a fitness ‘bonus’ due to it’s comparatively short length, and receive a score of 7.5. 24 A 300 character chromosome matching 5 lines of spam will be ‘punished’ for its long length, and receive a score of 2.5. 25

\[
24\ 5 \times (1 + \frac{200-100}{200}) = 5 \times (1 + \frac{100}{200}) = 5 \times 1.5 = 7.5
\]

\[
25\ 5 \times (1 + \frac{200-300}{200}) = 5 \times (1 + \frac{-100}{200}) = 5 \times (1 - .5) = 5 \times .5 = 2.5
\]

Eric Conrad
Without the length 'bonus,' chromosomes will grow longer each generation (simply adding other winning genes via breeding). The fitness function is designed to reward shorter/more efficient chromosomes, and keep the length manageable.

**Roulette Wheel Selection**

Roulette Wheel selection is a method for choosing chromosomes from a breeding pool. Higher-scoring chromosomes are more likely to be chosen than lower-scoring: a chromosome with a score of 3 is three times more likely to be chosen than a chromosome scoring 1.

> An imaginary roulette wheel is constructed with a segment for each individual in the population... the size of the segment is based on the aptness of the particular individual. A fit individual will occupy a larger slice of the roulette wheel than a weaker one.

Here are the 5 highest-scoring chromosomes from the 1st generation of our results (see below). '/^[A-Z][a-z]+\s+Kingdom\./' scored 5.64, compared with '/Adjou\./', which scored 1.96. The latter has a proportionally larger 'slot' in the roulette wheel, and is therefore more likely to be chosen for breeding.
Detecting Spam with Genetic Regular Expressions

Crossover Function

Our crossover function creates 1 child from 2 parents, sharing genes from both.

Another approach is to create 2 children from 2 parents, each sharing a subset of the parent’s genes.

Each survivor from the previous generation creates a child chromosome. A 2\textsuperscript{nd} parent is chosen from the breeding pool via Roulette Wheel selection. We have 2 fundamental crossover options: breaking down and combining all genes via an ‘or’ function (the ‘Or

---


27 Example simplified with 5 chromosomes for clarity. A ‘full’ roulette wheel would include all surviving chromosomes in its breeding pool.

Eric Conrad
Detecting Spam with Genetic Regular Expressions

Method'), and concatenating 2 parent’s chromosomes (the ‘Cat Method’).

The ‘Or Method’

An example of the ‘Or Method’ takes 2 parent chromosomes, such as ‘(FOO|BAR)’ and ‘(BAR|BAZ)’, breaks them down to their unique genes, and ‘ors’ them together to create a child called ‘(FOO|BAR|BAZ)’.

Here’s a specific example, taken from the generation 2 output, below. Here are 2 parents:

/\(?:^DR\./\s+[A-Z][a-z]+\|s+[A-Z][a-z]+\|^[A-Z][a-z]+\|s+Kingdom\.)/\n/\(?:^London,England|^\[A-Z\][a-z]+\|s+Kingdom\.)/\

The ‘Or Method’ breaks them down to their unique genes:

- ^London,England
- ^[A-Z][a-z]+\|s+
- Kingdom\.
- ^DR\./\s+
- [A-Z][a-z]+

It then creates a child chromosome by combining them in ‘OR’ (‘|’) fashion:

/\(?:^London,England|^\[A-Z\][a-z]+\|s+\|Kingdom\.|^\DR\./\s+\|[A-Z][a-z]+\)/

Barring mutation (see below), this child is unlikely to survive. It includes the gene ‘[A-
Detecting spam with genetic regular expressions

\[ z[a-z]+', \] which will match any capitalized word. The nature of crossover is both weaker stronger children will be created, and the strongest will survive.

The Cat Method

The 'Cat Method' takes 2 parent chromosomes and concatenates them together, creating a longer chromosome. Take these 2 parents:

\[
/(:\text{Coulibaly}\.|\text{LOTTERY}\s+\text{WINNING}\s+|^\text{deposit}|\text{the}\s+\text{finance})/
\]

\[
/YOUR\s+[a-zA-Z]+\s+[A-Z]+/
\]

Simply concatenate them, taking (/Chromosome1/) and (/Chromosome2/), creating the child (/Chromosome1|Chromosome2/):

\[
/(:\text{Coulibaly}\.|\text{LOTTERY}\s+\text{WINNING}\s+|^\text{deposit}|\text{the}\s+\text{finance}\|\text{YOUR}\s+[a-zA-Z]+\s+[A-Z]+)/
\]

Mutation

The mutation function takes a child chromosome, and changes it in a randomized way. The simplest method is deleting a random amount of genes.

Our mutation function takes a child gene, such as the child from the last section:

\[
/(:\text{Coulibaly}\.|\text{LOTTERY}\s+\text{WINNING}\s+|^\text{deposit}|\text{the}\s+\text{finance}\|\text{YOUR}\s+[a-zA-Z]+\s+[A-Z]+)/
\]

It then randomly deletes some genes:
Detecting Spam with Genetic Regular Expressions

/(?:LOTTERY\s+WINNING\s+|^deposit|the\s+finance|YOUR\s+[A-Z]+\s+[A-Z]+)/

Since shorter chromosomes are considered more ‘fit’ than equivalent longer chromosomes, mutation may help the fitness of a child. Also, mutation may remove some harmful genes, such as ‘[A-Z][a-z]+’.

As with Crossover, some mutations are harmful to children, but some may be helpful to the fitness of the child.

A word on efficiency

Due to the large amount of regular expression checks required in this process, the checks should be as efficient as possible. This regular expression contains grouping:

/Read the Jargon File to learn about the metasyntactic variable (Foo|Bar|Baz)/

In addition to grouping ‘Foo’, ‘Bar’ and ‘Baz’, it also captures the resulting match. To check to see what was actually matched in Perl, refer to variable “$1”:

print “The previous regex matched $1
”;

Capturing takes extra time and memory, so it is wasteful if the results of the capturing are not needed. Insert “?:” after the opening parentheses In order to avoid capturing (and make the regex more efficient):

/This regex groups but does not capture: (?=Foo|Bar|Baz)/

Eric Conrad
Detecting Spam with Genetic Regular Expressions

The code used in this paper will use efficient constructs such as ‘?:’ Where appropriate. The proof-of-concept script uses other efficiency improvements, such as pre-compiling the regexes. See the POC code in Appendix B for more details.
Detecting Spam with Genetic Regular Expressions

5. Proof-of-Concept Script Results

Here are the results of ‘genregex.pl’ (see Appendix B for full source code), running for 10 generations. The top 5 chromosomes are shown for each generation.

Note: these results are summarized and formatted into tables for clarity. The full text output is shown in Appendix C.

Generation 1

<table>
<thead>
<tr>
<th>Score</th>
<th>Chromosome</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.64</td>
<td><code>/^[A-Z][a-z]+\s+Kingdom\./</code></td>
</tr>
<tr>
<td></td>
<td><code>/^[a-z]+\s+to\s+[A-Z][a-z]+\s+[a-z]+\s+as\s+Consignment\s+[a-z]+\s+.*[A-Z]+\s+COURIER\s+/</code></td>
</tr>
<tr>
<td>3.13</td>
<td><code>/ive/</code></td>
</tr>
<tr>
<td>1.98</td>
<td><code>/^UK\.\</code></td>
</tr>
<tr>
<td>1.96</td>
<td><code>/Adjou\.\</code></td>
</tr>
</tbody>
</table>

The first generation contains automatic chromosomes, with no breeding or mutation (yet). The fittest chromosome, `/^[A-Z][a-z]+\s+Kingdom\./`, matches spam lines beginning with “United Kingdom.” It matches these three lines in the spam corpus:

```
# perl -e 'while<>{print if (/^[A-Z][a-z]+\s+Kingdom\./);}' < 419.txt
United Kingdom.
United Kingdom.
United Kingdom. I am writing following an opportunity in my office that
```

28 “A line of text beginning with a capital letter, followed by 1 or more lower-case letters, followed by one or more whitespace characters, followed by ‘Kingdom.’”

Eric Conrad
Detecting Spam with Genetic Regular Expressions

Although it only matches 3 lines, it receives a fitness bonus due to its short length, for a total fitness score of 5.64. Note: to see the spam lines matched by each chromosome, use the following command:

```
perl -e 'while(<>) {print if (/<REGEX>/);} ' < 419.txt
```

This is a simple Perl script that takes standard input (the file 419.txt), checks each line against a regex, and prints the line of there is a match.

Generation 2

The 2nd generation contains some chromosomes that have been bred and/or mutated:

<table>
<thead>
<tr>
<th>Score</th>
<th>Chromosome</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.82</td>
<td>'/Kingdom./'</td>
</tr>
<tr>
<td>7.74</td>
<td>'/the\s+finance/'</td>
</tr>
<tr>
<td>7.34</td>
<td>'/(?::^ive</td>
</tr>
<tr>
<td>7.12</td>
<td>'/(?::^London,England</td>
</tr>
<tr>
<td>6.76</td>
<td>'/(?::^DR./\s+[A-Z][a-z]+\s+[A-Z][a-z]+</td>
</tr>
</tbody>
</table>

The chromosome '/Kingdom\./' wins this round. It is shorter (and therefore fitter) than '/A-Z[a-z]+\s+Kingdom\./', and also matches an extra line:

```
perl -e 'while(<>) {print if (/Kingdom\./);} ' < 419.txt
United Kingdom.
United Kingdom.
United Kingdom. I am writing following an opportunity in my office that
acted by wire transfer via our correspondent bank, Citibank United Kingdom.
```

We also see the first sign of bred chromosomes, showing genes from multiple parents:

Eric Conrad
Detecting Spam with Genetic Regular Expressions

/(?:^London, England|^[A-Z][a-z]+\s+Kingdom\.)/

That matches spam containing 'London, England', or 'United Kingdom.'

Generation 3

<table>
<thead>
<tr>
<th>Score</th>
<th>Chromosome</th>
</tr>
</thead>
<tbody>
<tr>
<td>9.93</td>
<td>/(?:Kingdom.</td>
</tr>
<tr>
<td>9.57</td>
<td>/(?:or|s+his|s+.*of|s+kin.</td>
</tr>
<tr>
<td>9.47</td>
<td>/(?:Adjou.</td>
</tr>
<tr>
<td>9.38</td>
<td>/(?:Diara.</td>
</tr>
<tr>
<td>8.88</td>
<td>/(?:Coulibaly.</td>
</tr>
</tbody>
</table>

Bred chromosomes now hold the top 5 scores.

The top-scoring chromosome in generation 3 matches the following spam text:

```
$ perl -e 'while(<>) {print if/(?:Kingdom\.|Johnson\.|Mba\.|^PlasmaNet\.+\.*PO\|s+Box\|s+.\*> Station,\|s+\)/;}' < 419.txt
Barrister Johnson Mba.
PlasmaNet Inc. PO Box 4562, Grand Central Station, New
United Kingdom.
United Kingdom.
I am writing following an opportunity in my office that
tected by wire transfer via our correspondent bank, Citibank United Kingdom.
```

Generation 4

<table>
<thead>
<tr>
<th>Score</th>
<th>Chromosome</th>
</tr>
</thead>
<tbody>
<tr>
<td>25.72</td>
<td>/(?:LOTTERY|s+</td>
</tr>
<tr>
<td>20.05</td>
<td>/(?:Kingdom.</td>
</tr>
<tr>
<td>15.85</td>
<td>/(?:|s+|ive|^deposit)/</td>
</tr>
<tr>
<td>15.35</td>
<td>/(?:the|s+finance</td>
</tr>
<tr>
<td>15.35</td>
<td>/(?:Coulibaly.</td>
</tr>
<tr>
<td>13.8</td>
<td>/(?:MANUEL</td>
</tr>
</tbody>
</table>
Detecting Spam with Genetic Regular Expressions

The grouped chromosome:

`/(?:LOTTERY\s+|Lottery\s+|WINNING\s+|Coulibaly\.|^deposit)/`

...jumps out to the lead, with a score over 2 ½ times higher than the previous generation’s best. This chromosome shows characteristics of many high-scoring chromosomes: multiple short genes that each match large amounts of spam, in `(gene1|gene2|gene3...)` format.

Generation 5

<table>
<thead>
<tr>
<th>Score</th>
<th>Chromosome</th>
</tr>
</thead>
<tbody>
<tr>
<td>27.65</td>
<td>`/(?:MANUEL</td>
</tr>
<tr>
<td>25.72</td>
<td>`/(?:LOTTERY\s+</td>
</tr>
<tr>
<td>25.72</td>
<td>`/(?:Kingdom.</td>
</tr>
<tr>
<td>19.04</td>
<td>`MANUEL</td>
</tr>
<tr>
<td>16.17</td>
<td>`NIGERIA.</td>
</tr>
<tr>
<td>15.85</td>
<td>`\s+</td>
</tr>
</tbody>
</table>

In generation 5, a new Chromosome takes the lead. In addition to the ‘Lottery’
Detecting Spam with Genetic Regular Expressions

matches, it adds genes matching the names of alleged widows, such as 'Lorita Manuel,' and
'Mary Coulibaly.'

Generation 6

<table>
<thead>
<tr>
<th>Score</th>
<th>Chromosome</th>
</tr>
</thead>
<tbody>
<tr>
<td>29.64</td>
<td>/(?:^MISS\s+</td>
</tr>
<tr>
<td>27.82</td>
<td>/(?:LOTTERY\s+</td>
</tr>
<tr>
<td>27.65</td>
<td>/(?:MANUEL</td>
</tr>
<tr>
<td>26.62</td>
<td>^deposit)/</td>
</tr>
<tr>
<td>25.72</td>
<td>/(?:LOTTERY\s+</td>
</tr>
</tbody>
</table>

Generation 5's winner survives with the lead in generation 6. Three new chromosomes take slots 2-4.
Detecting Spam with Genetic Regular Expressions

Generation 7

<table>
<thead>
<tr>
<th>Score</th>
<th>Chromosome</th>
</tr>
</thead>
<tbody>
<tr>
<td>29.64</td>
<td>/(?::MISS\s+</td>
</tr>
<tr>
<td>29.44</td>
<td>/(?::MISS\s+</td>
</tr>
<tr>
<td>27.82</td>
<td>/(?::MISS\s+</td>
</tr>
<tr>
<td>27.65</td>
<td>/(?::MANUEL</td>
</tr>
<tr>
<td>27.27</td>
<td>/(?::MANUEL</td>
</tr>
</tbody>
</table>

Once again, generation 5’s winner holds on to the lead in generation 7. A new chromosome takes the #2 slot.

Generation 8

<table>
<thead>
<tr>
<th>Score</th>
<th>Chromosome</th>
</tr>
</thead>
<tbody>
<tr>
<td>30.24</td>
<td>/YOUR\s+[A-Z]+\s+[A-Z]+/</td>
</tr>
<tr>
<td>29.64</td>
<td>/(?::MISS\s+</td>
</tr>
<tr>
<td>29.44</td>
<td>/(?::MISS\s+</td>
</tr>
<tr>
<td>27.82</td>
<td>/(?::MANUEL</td>
</tr>
<tr>
<td>27.65</td>
<td>/(?::MANUEL</td>
</tr>
</tbody>
</table>

A new chromosome takes the lead, an automatic chromosome matching capital ‘YOUR’, followed by two capitalized words (each word followed by at least one whitespace character.

Eric Conrad
Detecting Spam with Genetic Regular Expressions

This chromosome receives a high fitness score because it is short, and it matches emails written in ALL CAPS (as many 419 scams are, to imply urgency), asking 'YOU' to do something.

The following spam lines are matched:

```
# perl -e 'while(<>){print if (/\(?:\text{Coulibaly\.|LOTTERY}\$+\text{WINNING}\$+\|^\text{deposit}\$+\text{the}\$+\text{finance}\$/
 \(?:\text{YOUR}\$+[A-Z]+\$+[A-Z]+\$/
 \(?:\text{YOUR}\$+[A-Z]+\$+[A-Z]+\|^\text{ive}\$+[A-Z]+\$+[a-z]+\$+[s+\text{Kingdom}\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$.\$
Detecting Spam with Genetic Regular Expressions

/\(?:\^MISS\s+|^ive|^MANUEL|^LOTTERY\s+|^Lottery\s+|^WINNING\s+|^LORITA\s+|^Coulibaly\./|^deposit|^the\s+finance|^Coulibaly\./|^deposit)/\n
29.44 \ LOTTERY\s+WINNING\s+|^deposit)/\n
Last generation’s winner is bred with another strong chromosome, chosen via Roulette Wheel selection, resulting in a chromosome that matches the ‘YOUR..’ capitalized text, ‘Mrs’ Coulibaly, the phrase ‘LOTTERY WINNING’, lines beginning with ‘deposit’ and the phrase ‘the finance’.

Here are the lines of spam generation 9’s winning chromosome matches:
Detecting Spam with Genetic Regular Expressions

### Generation 10

<table>
<thead>
<tr>
<th>Score</th>
<th>Chromosome</th>
</tr>
</thead>
<tbody>
<tr>
<td>50.58</td>
<td>`/(?::Coulibaly.</td>
</tr>
<tr>
<td>39.75</td>
<td>`/(?::Coulibaly.</td>
</tr>
</tbody>
</table>
| 31.17 | `/YOUR\s+[A-Z]+\s+[A-Z]+\s+\s+\/^MISS\s+|^ive|MANUEL|LOTTERY\s+|^winning\s+|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^ive|^i
Detecting Spam with Genetic Regular Expressions

In the final generation, a new chromosome takes the lead, scoring over 50. It combines the genes from generation 9’s winner, adding the gene ’MANUEL’ (matching lines containing ‘MISS LORITA MANUEL’, and the gene ‘Lottery\s+’ (matching any line containing ‘Lottery,’ followed by whitespace.

Here are the results from the winner of the final generation:
Detecting Spam with Genetic Regular Expressions

Summary

Here is a summary graph, showing the performance of the top-10 scoring chromosomes, over 10

Eric Conrad
Detecting Spam with Genetic Regular Expressions

Sample SpamAssassin Rule

This SpamAssassin rule is based on the winning genregex.pl rule:

```
body GENETIC_REGEX_1          /(?:YOUR\s+[A-Z]+\s+[A-Z]+|^MISS\s+|^ive|MANUEL|LOTTERY\s+|Lottery\s+|WINNING\s+|LORITA\s+|Coulibaly\.|^deposit)/

score GENETIC_REGEX_1          1.0

describe GENETIC_REGEX_1       'Advanced Fee' spam rule written by genregex.pl
```

Eric Conrad
Detecting Spam with Genetic Regular Expressions

Efficiency of genetic regular expressions

Size of the corpora

The proof-of-concept example run described above was run on a FreeBSD 6.2 server with a 3.2 gigahertz Xeon processor, and 2 gigs of memory.

The spam corpus contained 1155 unique lines, and the ham corpus contained 14799.

The run described in this paper generated the following statistics:

Comparisons: 8088. Comparisons/sec:38.8846153846154. Time taken was 208 seconds

208 seconds is a reasonable amount of time for a proof-of-concept script, but much larger corpora are required for truly useful operational results. The ham corpora did not contain the word ‘Lottery,’ for example, but it’s a word that does exist in non-spam email.

Much larger corpora (10 times+) would be useful for an operational environment (avoiding words like ‘Lottery’), which would result in program execution time of hours or more.

A note on ‘ors’

Large amounts of ‘ors’ (‘|’) become inefficient, as the regex engine must rescan a given multiple times for each ‘or’.

Eric Conrad
Detecting Spam with Genetic Regular Expressions

This is the winning chromosome:

```
/(?:<YOUR\s+[A-Z]+\s+[A-Z]+|^MISS\s+|^ive|MANUEL|
LOTTERY\s+|Lottery\s+|WINNING\s+|LORITA\s+|
Coulibaly\.|^deposit)/
```

The regex engine must take a line of text, and run the following (simplified) checks for the winning chromosome:

1. Scan the entire line for `YOUR\s+[A-Z]+\s+[A-Z]+`
2. Scan the beginning of the line (only) for `MISS\s+`
3. Scan the beginning of the line (only) for `ive`
4. Scan the entire line for `MANUEL`
5. Scan the entire line for `LOTTERY\s+`
6. Scan the entire line for `Lottery\s+`
7. Scan the entire line for `WINNING\s+`
8. Scan the entire line for `LORITA\s+`
9. Scan the entire line for `Coulibaly\.`
10. Scan the beginning of the line (only) for `deposit`

‘Anchored’ searches (using ‘^’, for example) are faster than unanchored searches, which must search the entire line for matching text. In the example run, the fittest chromosomes contained lots of ‘or’ operations, resulting in slower matches.

---

Eric Conrad
Detecting Spam with Genetic Regular Expressions

6. Conclusion

Genetic regular expressions effectively match spam, and show improvement generation-to-generation. The winning chromosome from the 10th generation was over 9 times as effective as the winning chromosome from the first.

Genetic Regular expressions exhibit the following characteristics, as described by Hiu Wong: ‘The basic techniques of the GAs are designed to simulate processes in natural systems necessary for evolution, specially those follow the principles first laid down by Charles Darwin of "survival of the fittest."’

Genetic regular expressions leverage the Genetic Algorithm concepts of fitness, crossover, and mutation to evolve chromosomes across generations to find a superior solution.

____________________


Eric Conrad
7. Further study

**Evolving Regex Primitives**

A promising field of study is evolving regex primitives used by genetic regular expressions. Certain regex design decisions were made in this paper, such as using actual lines of spam as the basis for chromosomes, joining multiple genes with ‘or’ operations, ensuring grouping parentheses always match, etc. These design decisions were made due to their practical and efficient application towards the problem domain.

A ‘primitive approach’ would break regexes down further, creating a ‘primordial soup’ of characters, including all literal characters that may be found in spam, and all regex metacharacters. Regexes could be built from this ‘soup.’ Initial generations would mostly comprise of nonsense strings of illegal regexes (unbalanced parentheses, anchor metacharacters in the middle of a string, etc) which would quickly die.

The approach would require building a regex analyzer that could not only determine ‘legal/illegal’ status of regex syntax, but also assign a syntax score: gibberish would receive a low score, and a regex that is one parenthesis away from ‘legal’ would receive a higher score.

Illegal regexes would be allowed to survive in early generations (until sufficient legal
Detecting Spam with Genetic Regular Expressions

regexes out-survive them), to help get over the initial 'legal regex' survival threshold ('crawl before you walk'). Ham-matching regexes would also survive (for a limited number of generations), until they are 'beaten' by spam-matching regexes.

Here is the logic for such an approach

1. Create regex primordial soup
2. Loop until done:
   a. Generate regexes from the 'soup'
   b. Assign a regex syntax score (0 to 1, graded from gibberish to legal syntax)
   c. Check for ham matches (score 1 if match)
   d. Check and score for spam matches (score 2+ if match)
   e. Keep the top-scoring 3rd
   f. Create children based on any survivors
   g. Mutate children (deleting genes, or adding random genes from the soup)

It would probably take hundreds or more generations (representing days of 3 ghz Xeon-equivalent CPU time) to evolve useful regexes, but, by leveraging Darwinian survival mechanisms, the regexes themselves could prove more powerful that those generated in POC code in this paper.
Detecting Spam with Genetic Regular Expressions

8. Appendices

Proof-of-concept code, including the spam and ham corpora used in this paper, may be downloaded from: http://files.ericconrad.com/genregex.tgz

The author may be reached at eric@ericconrad.com

Appendix A  ‘autoregex.pl,’ POC Perl Script demonstrating Automatic Regular Expressions

#!/usr/bin/perl
my $regex="";
open(HAMTXT,"<hams.txt")||die;
@hamlines = <HAMTXT>;
close(HAMTXT);

open(SPAMTXT,"<spams.txt")||die;
@spamlines = <SPAMTXT>;
close(SPAMTXT);

foreach (@spamlines) {
    chomp;
    my $newstring="";
    if (/^	/) {
        $newstring="^\t";
    }
    $len=length($_);
    $count=0;
    $wordcount=0;
    @word="";
    @line=split(/\//);
    while ($count<$len) {
        # Break string into a series of tokens
        # each word is a token, and each symbol (like "@" or "]") is a token
        # ":" is a token, unless it's in the format /^\w:\s/ (Matches headers like ^Subject: /)
        if (($line[$count] =~ /[A-Za-z0-9]/) || (($line[$count] =~ /*/) && ($line[$count+1] =~ /\s/))) {
            $wordcount++ if (($count!=0) && ($line[$count-1] =~ /[A-Za-z0-9]/));
            $word[$wordcount] .= $line[$count];
        }
    }
}

Eric Conrad
Detecting Spam with Genetic Regular Expressions

else{
    $wordcount++ if ($count!=0);
    # Escape all metacharacters we want to treat as literals
    if ($line[$count]=~/\([{}\[\]()^$\.]\+\?\|\)/){
        $word[$wordcount]="\"
    }
    $word[$wordcount].=$line[$count];
    $count++;
}

$first=1;
foreach (@word){
    $wordlen=length($_);
    if ($count!=0){
        # Match strings like /^Subject: /
        $newstring="^$_
        $first=0;
    }
    elsif ($wordlen>1){
        # If token is a word (>1 char)
        if (/\d+/){
            $newstring.="\d+
        if (/^[\w]+$/){
            $newstring.="^[\w]+:$/
        }
    }
    elsif (/^[A-F0-9]+$/){
        $newstring.="[A-F0-9]+";
    }
    elsif (/^[a-f0-9]+$/){
        $newstring.="[a-f0-9]+";
    }
    elsif (/^(com|net|org|edu|biz|info|us)$/){
        $newstring.="(com|net|org|edu|biz|info|us)";
    }
    elseif (/^[a-z]+$/){
        $newstring.="[a-z]+";
    }
    elsif (/^[A-Z]+$/){
        $newstring.="[A-Z]+";
    }
    elsif (/^(Mon|Tue|Wed|Thu|Fri|Sat|Sun)$/){
        $newstring.="(Mon|Tue|Wed|Thu|Fri|Sat|Sun)";
    }
    elsif (/^(Jan|Feb|Mar|Apr|May|Jun|Jul|Aug|Sep|Oct|Nov|Dec)$/){
        $newstring.="(Jan|Feb|Mar|Apr|May|Jun|Jul|Aug|Sep|Oct|Nov|Dec)";
    }
    elseif (/^[A-Z][a-z]+$/){
        $newstring.="[A-Z][a-z]+";
    }
    else{
        $newstring=
    }

Eric Conrad
Detecting Spam with Genetic Regular Expressions

```
$newstring.="$_";
}
else{  # Token is 1 character
  if (/[0-9]/){
    $newstring.="[0-9]";
  }
  elsif (/[A-Z]/){
    $newstring.="[A-Z]";
  }
  elsif (/[a-z]/){
    $newstring.="[a-z]";
  }
  elsif (/	/){
    $newstring.="\t";
  }else {
    $newstring.="$_";
  }
}
if ($newstring =~ /  +/){
  #print "$newstring\n";
  $newstring =~ s/  +/  \+/g;
  #print "$newstring\n";
}
$candidates{$_}=1;
}

# Delete any regext that match ham
while ( ($regex, $count) = each %candidates ) {
  foreach(@hamlines){
    if ((/$regex/g)&&(!$hamhit{$regex})){
      $hamhit{$regex}=1;
      #print "MISS: $regex\n";
    }
  }
}

# Rank remaining regex by # of spams matched
while ( ($regex, $count) = each %candidates ) {
  foreach(@spamlines){
    if ((/$regex/g)&&(!$hamhit{$regex})){
      $regexes{$regex}++;
      #print "HIT: $regex\n";
    }
  }
}
@sorted = sort { $regexes{$a} <=> $regexes{$b} } keys %regexes;
foreach (@sorted){
}

```

Eric Conrad

48
Detecting Spam with Genetic Regular Expressions

Appendix B: 'genregex.pl', Proof-of-concept Perl Script demonstrating Genetic Regular Expressions

#!/usr/bin/perl -w

# genregex.pl  v. 0.1
# Eric Conrad, September 2007
# genregex@ericconrad.com

# This program is free software, licensed under the GNU GPL, >=2.0.
# This software comes with absolutely NO WARRANTY. Use at your own risk!

use strict;

my %regexes;         # Hash that tracks all regexes
my %scores;          # Hash that tracks scores of chromosomes
my @fittest;         # Array containing top third chromosomes, sorted by score

my $start = time();  # Start of program execution
my @hamlines;        # array containing ham text
my @spamlines;       # array containing spam text
my $generations=10;  # Number of generations to evolve
my $generation=0;    # The current generation
my %hamhit;          # Hash that tracks whether a regex matched ham
my @survivors = ();  # The survivors chromosomes AoA
my %survivorhash;    # hash of references to the survivors AoA, key is the regex itself
my $score=0;         # Weighted score of a regex (matches * multiplier)
my $comparisons=0;   # Used to track # of regex comparisons

# Program logic:
#
# Read the spam and ham corpora
# Divide spam corpus into 10 slices
# Loop for 10 generations:
#   Create new chromosomes based on the next slice of spam
#   Create regex strings based on the chromosomes
#   Rank the chromosomes
#   Keep 1/3rd based on the highest score
#   Create a child for each of the current third

readcorpus(@hamlines,@spamlines);    # Read the spam and ham corpora
fisher_yates_shuffle(@spamlines);    # Shuffle the spam lines randomly
Detecting Spam with Genetic Regular Expressions

my $spamlines=scalar(@spamlines);       # Number of lines of spam
my $fract=int($spamlines/$generations); # Percentage of spam to process each
generation
my $thirtieth=int($spamlines/30);       # 10 generations, 1/3 of each is a
thirtieth

# Main program loop
while($generation <$generations ){
    print "Generation $generation ######################\n";
    my $offset=int($generation*$fract);   # Offset to gauge which chunk of spam is
    being processed
    create_chromosomes($offset,($offset+$fract),@survivors);  # Create new
    chromosomes based off the spam
    %survivorhash=stringify(@survivors); # Create regex strings based on @survivors
    misses(@hamlines);        # Ignore any survivors that match ham
    %regexes=hits(@spamlines);            # Tabulate hits per survivors
    %scores=scoreregexes(%regexes);       # Reward shorter survivors (and punish
    longer ones)
    @fittest=fittest(%scores);           # Grab the top third scores
    @survivors=breed(@fittest);          # Breed chromosomes based on the top third
    $generation++;
}

# end timer
my $end = time();

# report
print "Comparisons: $comparisons. Comparisons/sec:" , ($comparisons/($end -
$start)), ", Time taken was " , ($end - $start), " seconds";

sub breed{
    my $fref=shift;
    my @fittest=$fref;
    my @survivors;
    my @child;
    my $counter=0;       #temporary counter
    my @breedingpool=();
    my $breedref;
    my $cutoff=10;
    my $fitcounter=0;

    # Populate the breeding pool. Higher scoring regexes get more chances to breed
    # Use 'Roulette wheel' selection, based on the regex score. A score of 1 gives
    # 1 slot in the breeding pool, score of 3 gives 3 slots, etc.

    # A slot in the pool is a pointer to the survivorhash. A regex with a score of
    10
    # gets 10 pointers.

    foreach (@fittest){
        $counter=$fitcounter;
Detecting Spam with Genetic Regular Expressions

```perl
$fitcounter+=int($scores{$_});
while($counter<$fitcounter){
    $breedingpool[$counter]=@{$survivorhash{$_}};
    $counter++;
}

$counter=0;
foreach (@fittest){
    printf("%2.2f /%s/\n",$scores{$_},$_) if ($counter < $cutoff);
    $counter++;
    my $index = int (rand @breedingpool);  # Select a chromosome from the breeding pool
    $breedref = $breedingpool[$index];
    @child=crossover(@{$survivorhash{$_}},$breedref);     # Create a child chromosome based on the current survivor
    optimize(@child,0);                                   # Optimize the chromosome.
    push @survivors, [ @child ];                         # Save the child chromosome
    push @survivors, [ @{$survivorhash{$_}} ];          # Also save the current survivor
}
return(@survivors);
}

sub fittest{
    my @sorted;
    my $sortlines;
    %scores=(@_);
    @sorted = sort { $scores{$b} <=> $scores{$a} } keys %scores;
    splice(@sorted,$thirtieth);       # Keep the best-scoring third
    return(@sorted);
}

sub scoreregexes{
    my %regexes=@_;
    my $length=0;
    my $multiplier=0;
    my $regex;
    my $matches;
    while(($regex, $matches) = each(%regexes)) {
        $length=length($regex);
        #$length=length($regex)/2;
        $multiplier=(1+((200-$length)/200));
        #$multiplier=1 if ($length>=100);
        $score=$multiplier * $matches;
        $scores{$regex}=$score;
    }
    return(%scores);
}

sub stringify{

Eric Conrad
```
Detecting Spam with Genetic Regular Expressions

Eric Conrad

# Generate atring regex based on a chromosome
my $sref=shift;
my @survivors= @{$sref};
my $cref; # Reference to a chromosome
my $regex="";
my %survivorhash = ();

for $cref (@survivors) {
    $regex="";
    foreach @$cref{
        $regex.= "$_"; # Create a regex string based on the chromosome
    }
    $survivorhash{$regex}=$cref; # Build hash of references to the survivors array of arrays, key is the regex itself
}
return(%survivorhash);
}

sub crossover{
    my $motherref = shift ( @_ );
    my $fatherref = shift ( @_ );

    my @mother = @{$motherref};
    my @father = @{$fatherref};

    my $mlength = scalar( @mother ); # Length of mother
    my $flength = scalar( @father ); # Length of father

    my @child;
    my $clength;

    if (int(rand(2))==1){ # 50% chance 2 chromosomes are broken down to their genes, and combined in (a | y | b | z ) format
        push @mother, @father; # Going to swap between mother and child for some array manipulations
        @child = grep( !/\)|\(|\||\./, @mother ); # Remove genes with (,), |, { or .*.
        # Grab unique genes and sort them. Sort avoids (a | b) and (b | a) appearing to be 2 different genes
        my %hash = map { $_, 1 } @child;
        @child = sort { $hash{ $a } cmp $hash{ $b } } keys %hash;
        $clength = scalar( @child );
        if ($clength>1){
            #mutate( @child ) if (int(rand(2))==1);
            mutate( @child );
            $clength = scalar( @child ) - 1; # Length may have changed
            while ($clength>0){
                splice ( @child, $clength, 0, "|" );
                $clength--;
            }
        }
        unshift @child, "(?:";
Detecting Spam with Genetic Regular Expressions

push @child,")"; }
}
else{

  # 4 cases of mutation:
  # 1. mother + father
  # 2. (?:mother)+ father
  # 3. mother + (?:father)
  # 4. (?:mother) + (?:father)

  # There is probably simpler/cleaner logic to perform the following. The goal
  # is to avoid costly nested groups,
  # and keep things in "(foo|bar|baz)" format, and not "((foo|bar)|baz)" format

  # In mutation occurs via deleting a random amount of genes. Depending on the
  # grouping characteristics, a father, mother, or both
  # may have genes deleted before they are bred. If oth parents are pre-grouped,
  # no deletion occurs

  # TODO: write a more robust deletion/mutation subroutine that handles all
  # cases, and mutates all types of children
  # as opposed to pre-mutation some parents only

  if (($father[$flength-1] eq "")&& ($mother[$mlength-1] eq "))){  # both
    mother and father are groups
    pop @mother;            # Remove the trailing "")"
    push @mother,"|";     # Remove the trailing ")"
    shift @father;         # Remove the leading "(?:"
    push @mother,@father;
    removeduplicates(@mother);
  }
  elsif ($mother[$mlength-1] eq ")"){ # mother is already grouped, father is not
    deletegenes(@father);
    pop @mother;            # Remove the trailing "")"
    push @mother,"|";     # Remove the trailing ")"
    push @mother,@father;
    push @mother,");
    removeduplicates(@mother);
  }
  elsif ($father[$flength-1] eq ")"){ # father is grouped, mother is not
    deletegenes(@mother);
    unshift @mother,"(?:"
    shift @father;        # Remove the leading "(?:"
    push @mother,@father;
    removeduplicates(@mother);
  }
  else{                     # Neither is grouped
    deletegenes(@mother);
    deletegenes(@father);
Detecting Spam with Genetic Regular Expressions

```
unshift @mother,"(?:";
push @mother,"|";
push @mother,@father;
push @mother,"\")";
}
@child=@mother;
return(@child);
}

sub deletegenes{
    # TODO: Combine with mutate(), below
    my $arrayref=shift;
    my @chromosome=$($arrayref);
    my $length=scalar(@chromosome);
    my $random=int(rand($length)); # Random integer, keyed off parent1 length
    my $amount=int(rand($length-$random));
    splice(@chromosome,$random,$amount);  # delete a random number of genes
    splice(@chromosome,$random,0,".*") if (($random>0)&&($random<$length));
}

sub mutate{
    # TODO: Combine with deletegenes(), above
    my $childref=shift;
    my @chromosome=$($childref);
    my $mutations=4; # Number of different mutations available
    my $amount=0;                  # Amount of genes to delete
    my $choice;
    my $length2=scalar(@chromosome);
    my $random=int(rand($length2)); # Random integer, keyed off parent1 length
    fisher_yates_shuffle(@chromosome);
    $amount=int(rand($length2-$random));
    splice(@chromosome,$random,$amount);  # delete a random number of genes
}

sub fisher_yates_shuffle {
    # Code from the Perl FAQ 'How do I suffle an array randomly?'
    # http://perldoc.perl.org/perlfaq4.html#How-do-I-shuffle-an-array-randomly%3f
    my $deck = shift;  # $deck is a reference to an array
    my $i = @$deck;
    while (--$i) {
        my $j = int rand ($i+1);
        @$deck[$i,$j] = @$deck[$j,$i];
    }
}

sub genregex{
    my $linesref=shift(@_);
    my @currentlines=@{$linesref};
    ```
Detecting Spam with Genetic Regular Expressions

my @chromosomes;  # Array of chromosome arrays
my @chromosome;   # One chromosome
my @spamline;     # Line of spam text

foreach (@currentlines) {
  # print "GEN: $_\n";
  chomp;
  s/\s+/ /g;  # Reduce repeated whitespace to 1 space
  @spamline = split(" ");  # Split words into an array
  @chromosome = create_genes(@spamline);  # Create chromosomes based on each word
  optimize(@chromosome, 1);  # Optimize the chromosome. Pass by reference; chromosome will be altered
  push @chromosomes, [ @chromosome ];
}
return (@chromosomes);

sub optimize {
  # Optimize the chromosomes by grouping repeated genes, and deleting redundant genes (like .\.*)
  # Altering the passed array, so all array operations must be made by reference

  my $repeats = 0;  # Number of repeated genes (to be grouped)
  my $baseregex = "";  # Regex before grouping

  my $cref = shift(@_);
  my $first = shift(@_);

  my $wordcount = (scalar @{$cref}) - 1;  # Number of words in the array. array begins at offset 0, not 1
  my $current = $wordcount;  # Current position of genes array

  # Take 2 passes; one to add whitespace and delete repeated ".\.*"; the 2nd to combine identical
  # repeated genes. Could probably be done more efficiently in one pass.

  # 1st pass
  while ($current > 0) {
    if (($first) && ($current != $wordcount)) {  # Append whitespace unless it's the last genes or ".\.*"
      $cref->[0] .= " \s+" unless ($cref->[0] eq ".\.*");
    }
    # Remove repeated ".\.*"
    while (($current) && ($cref->[0] eq ".\.*") && ($cref->[0] eq "\s+")) {
      splice (@$cref, $current, 1);
      $current--;
    }
  }
  $wordcount = (scalar @{$cref}) - 1;  # may have changed due to splice

Eric Conrad 55
Detecting Spam with Genetic Regular Expressions

```perl
$current=$wordcount;
# 2nd pass
while($current>0){
    $baseregex=$cref->[0];
    $repeats=0;
    # If there are repeated genes, combine them in (genes){count} format
    while (((($current) && ($cref->[0] eq $cref->[0]))){
        $repeats++;
        splice(@{$cref},$current,1);
        $current--;
    } if ($repeats){
        $repeats++;
        $cref->[0]="(?:$baseregex){$repeats}"; # ?: is ugly, but avoids capturing, which is slow
    }
    $current--;
}
if ($cref->[0] eq ".*"){ # Remove leading ".*"
    splice(@{$cref},0,1)
} else{
    # Append a "^" on the first run, and skip regexes that began with ".*"
    $cref->[0]="^$cref->[0]" if ($first);
}
$wordcount= (scalar @{$cref})-1; # may have changed due to splice
# Remove trailing ".*" (redundant). Check now, after multiple contiguous ".*"s have been consolidated
if ($cref->[0] eq ".*"){
    splice(@{$cref},$wordcount,1) if (($cref->[0]) eq ".*");  
}
}
sub removeduplicates{
    # combines duplicate genes, changing:
    # /(?FOO|You ([a-z]+\s+)2 eaten by a grue|FOO|BAR)/
    # to: /(?FOO|You (\[a-z]+\s+)(2 eaten by a grue|BAR)/
    # Simple shortcuts like creating a unique hash don't help, as 'eaten by a grue' is 4 genes that must remain in order
    # Logic is 2 loops: start with gene1, and search for dups in gene1+X
    # Then: gene2, and search for dups in gene2+X
    # Etc.
    # If a dup is found, delete the 2nd matching gene, plus the previous "|".
    # There is probably a cleaner way to do this
    #
    # TODO: Some dups are still slipping through, probably due to a bug in the below code.
    #
    #}
```

Eric Conrad
Detecting Spam with Genetic Regular Expressions

my $cref = shift(@_);
my @chromosome= @{$cref};
my $count=1;
my $length=scalar (@chromosome);
my @pipes=grep ("\|", @chromosome);
if ((@pipes) && (@chromosome[0] eq "(?:))"){
    while($count<$length){
        my $first=@chromosome[$count];
        $count++;
        my $count2=$count;
        while($count2<$length){
            $count2++;
            if (($count2<$length) && (@chromosome[$count2] eq "|")){
                $count2++;
                if (($count2<$length) && ($first eq $chromosome[$count2])){
                    my $offset=$count2-1;
                    $count2++;
                    if (($count2<$length) && (($chromosome[$count2] eq "|"))){
                        splice(@chromosome,$offset,2);
                        $length=scalar (@chromosome);
                    }
                }
            }
        }
    }
}

sub create_genes { # TODO: code creates some infrequent 'strict' warnings, need to be fixed
    my @spamline = @_; # Number of words in a line of spam
    my $wordcount= scalar @spamline;
    my $wordlen=0;
    my $choice=0;
    my $length=0;
    my @chromosome=();
    foreach (@spamline){
        s/'([^{}\[\]\(\)\^\$\.\|\*\+\?\\])/\$1/g; # Escape all metacharacters
        $wordlen=length($_);
        $choice=int(rand(3));
        if ($choice == 0){
            push @chromosome, ".*";
        } elsif ($choice < ($wordcount/4)){
            push @chromosome, ":";
        } elsif ($choice==1){
            push @chromosome, "_";
        } else{
            push @chromosome, "\$_";
        }
    }
}
Detecting Spam with Genetic Regular Expressions

```perl
if ($wordlen>1){    # If gene is a word (>1 char)
    if (/^(.)\1{2,}$/) {    # Repeated character
        $length=length($_);
        push @chromosome,"$_{$length}";
    }
    elsif (/\d+/){    # Numbers
        push @chromosome,"\d+";    # Numbers
    }
    elsif (/^[A-F0-9]+$/){    # Hex with caps
        push @chromosome,"[A-F0-9]{$wordlen}";
    }
    elsif (/^[a-f0-9]+$/){    # Hex with lowercase
        push @chromosome,"[a-f0-9]{$wordlen}";
    }
    elsif (/^(?:com|net|org|edu|biz|info|us)$/){    # lowercase letters
        push @chromosome,"(?:com|net|org|edu|biz|info|us)";
    }
    elsif (/^[a-z]+$/){    # lowercase letters
        push @chromosome,"[a-z]+";
    }
    elsif (/^[A-Z]+$/){    # Capital letters
        push @chromosome,"[A-Z]+";
    }
    elsif (/^(?:Mon|Tue|Wed|Thu|Fri|Sat|Sun)$/){    # Days of the week
        push @chromosome,"(?:Mon|Tue|Wed|Thu|Fri|Sat|Sun)";
    }
    elsif (/^(?:Jan|Feb|Mar|Apr|May|Jun|Jul|Aug|Sep|Oct|Nov|Dec)$/){    # Months of the year
        push @chromosome,"(?:Jan|Feb|Mar|Apr|May|Jun|Jul|Aug|Sep|Oct|Nov|Dec)";
    }
    elsif (/^[A-Z][a-z]+$/){    # Words With The First Letter Capitalized
        push @chromosome,"[A-Z][a-z]+";
    }
    elsif (/^(US)\$(0-9,)+$/){    # Cash amounts such as US$9,300,000
        #print "@spamline\n";
        push @chromosome,"(US)\$(0-9,)+";
        #print "@chromosome\n";
    }
    else{
        push @chromosome,"$_";
    }
}
else{    # Token is 1 character
    if (/^[0-9]$/){
        push @chromosome,"[0-9]";
    }
    elsif (/^[A-Z]$/){
        push @chromosome,"[A-Z]";
    }
    elsif (/^[a-z]$/){
        push @chromosome,"[a-z]";
    }
}
```

Eric Conrad
Detecting Spam with Genetic Regular Expressions

```
} 
else {
    push @chromosome,"$_";

}
}
return(@chromosome);
}

sub misses{
    my $hamref=shift;
    my @hamlines= @{$hamref};
    my $regex;
    my $compiled="";  # compiled regex
    while ( $regex = each %survivorhash ) {
        $compiled= qr/$regex/;  # Much more efficient to compile the regex in advance
        foreach(@hamlines){
            if (!($hamhit{$regex})){
                if (/$/compiled/) {
                    $comparisons++;
                    $hamhit{$regex}=1;
                    #print "MISS: $regex\n";
                }
            }
        }
    }
    #return($comparisons);
}

sub hits {
    my $spamref=shift;
    my @spamlines= @{$spamref};
    my $compiled="";  # compiled regex
    my %regexes;
    my $regex;
    while ( $regex = each %survivorhash ) {
        if ( !((!($regexes{$regex})||($hamhit{$regex}))) ) {  # If it's a new regex (no previous hits or misses)
            $compiled= qr/$regex/;
            foreach(@spamlines){
                if (/$/compiled/) {
                    $comparisons++;
                    $regexes{$regex}++;
                    #print "HIT: $regexes{$regex} $regex\n";
                }
            }
        }
    }
```

Eric Conrad

© SANS Institute 2007, As part of the Information Security Reading Room Author retains full rights.
Detecting Spam with Genetic Regular Expressions

sub readcorpus {
    # Read the spam and ham corpora from text files
    @hamlines=shift(@_);
    @spamlines=shift(@_);

    #open(HAMTXT,"<ham-bodies0.txt")||die;
    open(HAMTXT,"<ham.txt")||die;
    while(<HAMTXT>){
        push @hamlines,$_ if (/\w/);  # Save lines containing at least 1 word character
    }
    close(HAMTXT);

    #open(SPAMTXT,"<419-bodies0.txt")||die;
    open(SPAMTXT,"<419.txt")||die;
    while(<SPAMTXT>){
        push @spamlines,$_ if (/\w/);  # Save lines containing at least 1 word character
    }
    close(SPAMTXT);

    #return(@hamlines,@spamlines);
}

sub create_chromosomes{
    my $cref;
    my @currentlines;    # temp array containing spam text which may be spliced, etc.
    Needed to leave @spamlines intact
    my @currentchromosomes;   # AoA containing the current chromosomes in play
    my $offset=shift (@_);
    my $length=shift (@_);
    my $survivorref=shift (@_);
    # Seed the first generation
    @currentlines = @spamlines[$offset..$length];  # Grab a chunk of spam
    @currentchromosomes=genregex(@currentlines);  # Generate regexes based on the spam
    for $cref ( @currentchromosomes ) {
        push @{$survivorref}, [ @$cref ];      # Add the current chromosomes to the survivor array
    }
}

Appendix C, 'genregex.pl' output

# ./genregex.pl
Generation 0 ########################
5.64 ^[A-Z][a-z]+\s+Kingdom\./
Detecting Spam with Genetic Regular Expressions

3.13 /^[a-z]+\s+to\s+[A-Z]\[a-z]+\s+[a-z]+\s+as\s+Consignment\s+[a-z]+\s+.*\[A-Z]+\s+\s+CORI\s+/  
1.98 /ive/  
1.98 /UK\./  
1.96 /Adjou\./  
1.94 /Coulibaly\./  
1.93 /c\.Private\s+/  
1.93 /Dear\s+.\*One\./  
1.93 /[A-Z]+\s+CENTER/  
1.93 /London,\s+England/  
Generation 1  
7.82 /Kingdom\./  
7.74 /the\s+finance/  
7.34 /\^[a-z]+\s+[A-Z]\[a-z]+\s+\s+as\s+Consignment\s+[a-z]+\s+.*\[A-Z]+\s+\s+CORI\s+/  
7.12 /\^[a-z]+\s+to\s+[A-Z]\[a-z]+\s+as\s+Consignment\s+[a-z]+\s+.*\[A-Z]+\s+\s+CORI\s+/  
6.76 /\^[a-z]+\s+[A-Z]\[a-z]+\s+\s+as\s+Consignment\s+[a-z]+\s+.*\[A-Z]+\s+\s+CORI\s+/  
6.46 /\^[a-z]+\s+[A-Z]+\s+as\s+Consignment\s+[a-z]+\s+.*\[A-Z]+\s+\s+CORI\s+/  
Generation 2  
9.93 /\^[a-z]+\s+to\s+[A-Z]\[a-z]+\s+\s+as\s+Consignment\s+[a-z]+\s+.*\[A-Z]+\s+\s+CORI\s+/  
9.57 /\^[a-z]+\s+to\s+[A-Z]\[a-z]+\s+\s+as\s+Consignment\s+[a-z]+\s+.*\[A-Z]+\s+\s+CORI\s+/  
9.38 /\^[a-z]+\s+to\s+[A-Z]\[a-z]+\s+\s+as\s+Consignment\s+[a-z]+\s+.*\[A-Z]+\s+\s+CORI\s+/  
Generation 3  
25.72 /\^[a-z]+\s+to\s+[A-Z]\[a-z]+\s+\s+as\s+Consignment\s+[a-z]+\s+.*\[A-Z]+\s+\s+CORI\s+/  
15.85 /\^[a-z]+\s+to\s+[A-Z]\[a-z]+\s+\s+as\s+Consignment\s+[a-z]+\s+.*\[A-Z]+\s+\s+CORI\s+/  
15.35 /\^[a-z]+\s+to\s+[A-Z]\[a-z]+\s+\s+as\s+Consignment\s+[a-z]+\s+.*\[A-Z]+\s+\s+CORI\s+/  
13.80 /\^[a-z]+\s+to\s+[A-Z]\[a-z]+\s+\s+as\s+Consignment\s+[a-z]+\s+.*\[A-Z]+\s+\s+CORI\s+/  
12.69 /\^[a-z]+\s+to\s+[A-Z]\[a-z]+\s+\s+as\s+Consignment\s+[a-z]+\s+.*\[A-Z]+\s+\s+CORI\s+/  
11.08 /\^[a-z]+\s+to\s+[A-Z]\[a-z]+\s+\s+as\s+Consignment\s+[a-z]+\s+.*\[A-Z]+\s+\s+CORI\s+/  

Eric Conrad
Detecting Spam with Genetic Regular Expressions

Eric Conrad
Detecting Spam with Genetic Regular Expressions

Eric Conrad
Detecting Spam with Genetic Regular Expressions

29.64 /(?::^MISS\s+|^ive|MANUEL|LOTTERY\s+|Lottery\s+|WINNING\s+|LORITA\s+|Coulibaly\.||^deposit)/
29.44 /(?::^MISS\s+|^ive|MANUEL|LOTTERY\s+|Lottery\s+|WINNING\s+|LORITA\s+|Coulibaly\.||^deposit|the\s+|finance|Coulibaly\.||LOTTERY\s+|WINNING\s+||^deposit)/
27.82 /(?::LOTTERY\s+|Lottery\s+|WINNING\s+|Coulibaly\.||^deposit|or\s+|his\s+.*of\s+|kin\.||^account\s+{(?:[a-z]+\s+)\{2\}.*away,\s+san|the\s+finance)/
27.65 /(?::MANUEL|^MISS\s+|^ive|LORITA\s+|Coulibaly\.||^deposit|LOTTERY\s+|Lottery\s+|WINNING\s+|Coulibaly\.||^deposit)/
27.27 /(?::the\s+|finance|Coulibaly\.||LOTTERY\s+|WINNING\s+||^deposit|LOTTERY\s+|Lottery\s+|WINNING\s+|Coulibaly\.||^deposit)/
27.03 /(?::MANUEL|^MISS\s+|^ive|LORITA\s+|Coulibaly\.||^deposit|LOTTERY\s+|Lottery\s+|WINNING\s+|Coulibaly\.||^deposit|the\s+|finance)/
26.62 /(?::or\s+|his\s+.*of\s+|kin\.||^account\s+{(?:[a-z]+\s+)\{2\}.*away,\s+san|the\s+finance|MANUEL|^MISS\s+|^ive|LORITA\s+|Coulibaly\.||^deposit|LOTTERY\s+|Lottery\s+|WINNING\s+|Coulibaly\.||^deposit)/
25.72 /(?::LOTTERY\s+|Lottery\s+|WINNING\s+|Coulibaly\.||^deposit)/
24.61 /(?::Coulibaly\.||LOTTERY\s+|WINNING\s+||^deposit|the\s+|finance|Coulibaly\.||LOTTERY\s+|WINNING\s+||^deposit|MISS\s+|^ive|MANUEL|LOTTERY\s+|Lottery\s+|WINNING\s+|LORITA\s+|Coulibaly\.||^deposit)/
Generation 8 ####################################
39.75 /(?::Coulibaly\.||LOTTERY\s+|WINNING\s+||^deposit|the\s+|finance|YOUR\s+[A-Z]+\s+[A-Z]+)/
31.17 /(?::YOUR\s+[A-Z]+\s+[A-Z]+|^ive|^A-Z^[A-Z]+|\s+Kingdom\.|Coulibaly\.|Johnson\s+Mba\.|^PlasmaNet\s+.*PO\s+Box\s+.*Station,\s+|^deposit|MANUEL|^MISS\s+|^ive|LORITA\s+|Coulibaly\.||^deposit)/
30.24 /YOUR\s+\s+[A-Z]+\s+[A-Z]+)/
29.64 /(?::^MISS\s+|^ive|MANUEL|LOTTERY\s+|Lottery\s+|WINNING\s+|LORITA\s+|Coulibaly\.||^deposit)/
29.44 /(?::^MISS\s+|^ive|MANUEL|LOTTERY\s+|Lottery\s+|WINNING\s+|LORITA\s+|Coulibaly\.||^deposit|the\s+|finance|Coulibaly\.||LOTTERY\s+|WINNING\s+||^deposit)/
27.82 /(?::LOTTERY\s+|Lottery\s+|WINNING\s+|Coulibaly\.||^deposit|or\s+|his\s+.*of\s+|kin\.||^account\s+{(?:[a-z]+\s+)\{2\}.*away,\s+san|the\s+finance)/
27.65 /(?::MANUEL|^MISS\s+|^ive|LORITA\s+|Coulibaly\.||^deposit|LOTTERY\s+|Lottery\s+|WINNING\s+|Coulibaly\.||^deposit)/
27.27 /(?::the\s+|finance|Coulibaly\.||LOTTERY\s+|WINNING\s+||^deposit|LOTTERY\s+|Lottery\s+|WINNING\s+|Coulibaly\.||^deposit)/
27.03 /(?::the\s+|finance|Coulibaly\.||LOTTERY\s+|WINNING\s+||^deposit|MANUEL|^MISS\s+|^ive|LO
Detecting Spam with Genetic Regular Expressions

RITA\s+|Coulibaly\.|\^deposit|LOTTERY\s+|Lottery\s+|WINNING\s+|Coulibaly\.|\^deposit/  
27.03  
/\(?:\MANUEL|\^MISS\s+|\^ive|LORITA\s+|Coulibaly\.|\^deposit|LOTTERY\s+|Lottery\s+|WINNING\s+|Coulibaly\.|\^deposit\)/  
Generation 9  
50.58 /\(?:\YOUR\s+[A-Z]+\s+[A-Z]+\)+\^MISS\s+|\^ive|\MANUEL|LOTTERY\s+|Lottery\s+|WINNING\s+|LORITA\s+|Coulibaly\.|\^deposit/  
39.75 /\(?:\Coulibaly\.|\^deposit|\the\s+finance\)/  
31.17 /\(?:\YOUR\s+[A-Z]+\s+[A-Z]+\)+\^ive|\^[A-Z]+[A-Z]+\s+\^deposit/  
29.64 /\(?:\^MISS\s+|\^ive|\MANUEL|LOTTERY\s+|Lottery\s+|WINNING\s+|LORITA\s+|Coulibaly\.|\^deposit/  
29.44 /\(?:\^MISS\s+|\^ive|\MANUEL|LOTTERY\s+|Lottery\s+|WINNING\s+|LORITA\s+|Coulibaly\.|\^deposit/  
27.82 /\(?:\^LO\REETY\s+\^WINNING\s+\|Coulibaly\.|\^deposit|\or\s+\^his\s+.\^o\^s\+\^k\^in\.|\^account\s+\(?:\^[a-z]+\s+\(2\)+\^away,\s+\^san\|\the\s+finance\)/  
27.65 /\(?:\MANUEL|\^MISS\s+|\^ive|LORITA\s+|Coulibaly\.|\^deposit\)/  
27.27 /\(?:\the\s+\^finance\)/  
27.03 /\(?:\\the\s+\^finance\)/  
Comparisons: 8088.  Comparisons/sec: 38.8846153846154. Time taken was 208 seconds
Detecting Spam with Genetic Regular Expressions

9. References

Documents


Eric Conrad 66
Detecting Spam with Genetic Regular Expressions

Tools

GNU sort. URL: http://www.gnu.org/software/coreutils/

Perl. URL: http://www.perl.org/

Perl Compatible Regular Expressions. URL: http://www.pcre.org/

Postfix. URL: http://www.postfix.org/

SpamAssassin. URL: http://spamassassin.apache.org
<table>
<thead>
<tr>
<th>Course</th>
<th>Location</th>
<th>Dates</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>SANS Las Vegas 2020</td>
<td>Las Vegas, NV</td>
<td>Jan 27, 2020 - Feb 01, 2020</td>
<td>Live Event</td>
</tr>
<tr>
<td>SANS Vienna January 2020</td>
<td>Vienna, AT</td>
<td>Jan 27, 2020 - Feb 01, 2020</td>
<td>Live Event</td>
</tr>
<tr>
<td>SANS Security East 2020</td>
<td>New Orleans, LA</td>
<td>Feb 01, 2020 - Feb 08, 2020</td>
<td>Live Event</td>
</tr>
<tr>
<td>SANS New York City Winter 2020</td>
<td>New York City, NY</td>
<td>Feb 10, 2020 - Feb 15, 2020</td>
<td>Live Event</td>
</tr>
<tr>
<td>SANS Northern VA - Fairfax 2020</td>
<td>Fairfax, VAUS</td>
<td>Feb 10, 2020 - Feb 15, 2020</td>
<td>Live Event</td>
</tr>
<tr>
<td>SANS Dubai February 2020</td>
<td>Dubai, AE</td>
<td>Feb 15, 2020 - Feb 20, 2020</td>
<td>Live Event</td>
</tr>
<tr>
<td>SANS Cairo February 2020</td>
<td>Cairo, EG</td>
<td>Feb 15, 2020 - Feb 20, 2020</td>
<td>Live Event</td>
</tr>
<tr>
<td>SANS San Diego 2020</td>
<td>San Diego, CAUS</td>
<td>Feb 17, 2020 - Feb 22, 2020</td>
<td>Live Event</td>
</tr>
<tr>
<td>SANS Scottsdale 2020</td>
<td>Scottsdale, AZUS</td>
<td>Feb 17, 2020 - Feb 22, 2020</td>
<td>Live Event</td>
</tr>
<tr>
<td>SANS Brussels February 2020</td>
<td>Brussels, BE</td>
<td>Feb 17, 2020 - Feb 22, 2020</td>
<td>Live Event</td>
</tr>
<tr>
<td>Open-Source Intelligence Summit &amp; Training 2020</td>
<td>Alexandria, VAUS</td>
<td>Feb 18, 2020 - Feb 24, 2020</td>
<td>Live Event</td>
</tr>
<tr>
<td>SANS Training at RSA Conference 2020</td>
<td>San Francisco, CAUS</td>
<td>Feb 23, 2020 - Feb 24, 2020</td>
<td>Live Event</td>
</tr>
<tr>
<td>SANS Jacksonville 2020</td>
<td>Jacksonville, FLUS</td>
<td>Feb 24, 2020 - Feb 29, 2020</td>
<td>Live Event</td>
</tr>
<tr>
<td>SANS Secure India 2020</td>
<td>Bangalore, IN</td>
<td>Feb 24, 2020 - Feb 29, 2020</td>
<td>Live Event</td>
</tr>
<tr>
<td>SANS Manchester February 2020</td>
<td>Manchester, GB</td>
<td>Feb 24, 2020 - Feb 29, 2020</td>
<td>Live Event</td>
</tr>
<tr>
<td>SANS Zurich February 2020</td>
<td>Zurich, CH</td>
<td>Feb 24, 2020 - Feb 29, 2020</td>
<td>Live Event</td>
</tr>
<tr>
<td>SANS Northern VA - Reston Spring 2020</td>
<td>Reston, VAUS</td>
<td>Mar 02, 2020 - Mar 07, 2020</td>
<td>Live Event</td>
</tr>
<tr>
<td>SANS Munich March 2020</td>
<td>Munich, DE</td>
<td>Mar 02, 2020 - Mar 07, 2020</td>
<td>Live Event</td>
</tr>
<tr>
<td>SANS Secure Japan 2020</td>
<td>Tokyo, JP</td>
<td>Mar 02, 2020 - Mar 14, 2020</td>
<td>Live Event</td>
</tr>
<tr>
<td>Blue Team Summit &amp; Training 2020</td>
<td>Louisville, KYUS</td>
<td>Mar 02, 2020 - Mar 09, 2020</td>
<td>Live Event</td>
</tr>
<tr>
<td>SANS St. Louis 2020</td>
<td>St. Louis, MOUS</td>
<td>Mar 08, 2020 - Mar 13, 2020</td>
<td>Live Event</td>
</tr>
<tr>
<td>SANS Dallas 2020</td>
<td>Dallas, TXUS</td>
<td>Mar 09, 2020 - Mar 14, 2020</td>
<td>Live Event</td>
</tr>
<tr>
<td>SANS Prague March 2020</td>
<td>Prague, CZ</td>
<td>Mar 09, 2020 - Mar 14, 2020</td>
<td>Live Event</td>
</tr>
<tr>
<td>Wild West Hackin Fest 2020</td>
<td>San Diego, CAUS</td>
<td>Mar 10, 2020 - Mar 11, 2020</td>
<td>Live Event</td>
</tr>
<tr>
<td>SANS Doha March 2020</td>
<td>Doha, QA</td>
<td>Mar 14, 2020 - Mar 19, 2020</td>
<td>Live Event</td>
</tr>
<tr>
<td>SANS Norfolk 2020</td>
<td>Norfolk, VAUS</td>
<td>Mar 16, 2020 - Mar 21, 2020</td>
<td>Live Event</td>
</tr>
<tr>
<td>SANS San Francisco East Bay 2020</td>
<td>OnlineCAUS</td>
<td>Jan 27, 2020 - Feb 01, 2020</td>
<td>Live Event</td>
</tr>
<tr>
<td>SANS OnDemand</td>
<td>Books &amp; MP3s OnlyUS</td>
<td>Anytime</td>
<td>Self Paced</td>
</tr>
</tbody>
</table>