A Comprehensive Measurement Study of Domain Generating Malware

Daniel Plohmann, Fraunhofer FKIE; Khaled Yakdan, University of Bonn; Michael Klatt, DomainTools; Johannes Bader; Elmar Gerhards-Padilla, Fraunhofer FKIE

Acadêmico: Paulo Henrique

Abstract

– *Botnets* modernas vem adotando algoritmos geradores de domínios (DGA).

– DGAs são importantes para melhorar a infraestrutura de uma *botnet* e para dificultar o *blacklisting* e tentativas de derrubar a rede.

– Esse artigo faz um estudo do panorama dos algoritmos analisando 43 famílias de *Malwares* baseados em DGAs.

– Apresenta um taxonomia para os DGAs.

– Reimplementação das famílias.

1. Introdução

– *Botnets* são redes de máquina infectadas por um *Malware* e remotamente controlada por um *botmaster* através de um *C&C communication*.

– ‘Defenders’ tentam acabar com uma botnet através do canal C&C, fazendo ataques conhecidos como *sinkholing.*

– Um DGA é usado para dinamicamente gerar um grande número de domínios randômicos e selecionar um pequeno grupo para realizar a comunicação C&C.

1. Introdução

– É proposto uma taxonomia para categorizar os aspectos principais e usá-la para descrever e comparar os DGAs estudados.

– Essas análises são baseadas em engenharias reversas desses algoritmos.

2. A DGA Taxonomy

2.1 – *Seed Source*

– Incluem parâmetros como constantes numéricas (tamanho do domínio ou raiz randômica) e *strings* (alfabeto ou conjunto de possíveis *TLDs*).

– *Time dependence* – usa o tempo como *seed* para calcular os domínios. (a data em uma resposta HTTP)

– Determinismo – usa dados não previsíveis mas públicos. (Twitter)

2. A DGA Taxonomy

2.2 – *Generation Schemes*

- *Arithmetic-based* – calcula a sequencia de valores diretos como ASCII.

- *Hash-based* – representação hexadecimal (MD5 e SHA256).

- *Wordlist-based* – concatena strings de uma ou mais listas.

- *Permutation-based* – calcula as palavras derivadas de um nome de domínio inicial.

2. A DGA Taxonomy

2.3 – DGAs types

- TDD-A (20)
- TID-A (16)
- TDD-W (3)
- TDD-H (2)
- TDN-A (1)
- TID-P (1)

3. DGA-Malware Dataset

3.1 – Identifying DGA-based Malware

– *Filtering*

– filtra a domínios benignos (*Alexa list*);

– filtra domínios conhecidos comparando os domínios de entrada com os atuais domínios enumerados (*malware analysis reports* e blogs);

– foi identificado 22 famílias, onde 9 eram apenas re-implementações.

3. DGA-Malware Dataset

3.1 – Identifying DGA-based Malware

– *Identification*

– identificar amostras geradas por DGAs conhecidas mas usam outras chaves e possíveis DGAs;

– utiliza um conjunto de expressões regulares para decidir rapidamente se coincide com a saída esperada de um DGA conhecido.

3. DGA-Malware Dataset

3.1 – *Identifying DGA-based Malware*

– *DGA enumeration*

– quando uma amostra é identificada como novo DGA, ela é verificada manualmente fazendo uma engenharia reversa.

– como entrada foi fornecido uma *sandbox* especial pela *Shadowserver Foundation* (com novos e antigos *malwares*).

3. DGA-Malware Dataset

3.2 – *Reimplementing DGAs*

– compara a saída com os domínios apresentados pelo respectivo *malware* quando executado.

– para cada família, a engenharia reversa tomou em média 1 dia.

– extração da raiz pelo binário.

– após a extração do binário: sem camadas de proteção, o que agilizou o processo.

4. Insights into the DGA Landscape

4.1 – *Domain Structure*

– Alfabeto
  – entre 9 e 36 caracteres.
  – 11 DGAs com o bug *off-by-one*.
  – *truncation* erros (Szribi trunca de 26 para 15).

4. Insights into the DGA Landscape

4.1 – *Domain Structure*

– *AGD length*

– de 4 a 47 caracteres (media de 9 a 16).

– 14 DGAs produzem domínios de mesmo tamanho.

4. Insights into the DGA Landscape

4.1 – *Domain Structure*

– Níveis de domínios

– 3 famílias criam domínios de terceiro nível, os outros geram apenas de segundo nível.

4. Insights into the DGA Landscape

4.2 – *Domain Validity Periods*

– 24 são dependentes do tempo, onde 21 geram domínios com validades separadas.

– *Matsnu*: gera 3 domínios, válido por 3 dias consecutivos cada (implica em 9 potenciais C&C domínios por vez). (exceções para *Pushdo* e *Suppobox*).

– 11 DGAs geram domínios válidos por 1 dia.

4. Insights into the DGA Landscape

4.3 – *Generation Schemes*

– *Arithmetic-based* são os mais comuns, 37. 26 computam o código ASCII diretamente para ser usado no domínio, e 11 usam índices de *arrays*.

– 3 DGAs são baseados em listas de palavras.

– *Matsnu* e *Gozi* combinam palavras até um tamanho determinado e adiciona *.net*.

4. Insights into the DGA Landscape

4.4 – *Domain Randomness*

– concatena todas as *strings* geradas pelo DGA e calcula a entropia de Shannon.

– calcula a entropia relativa.

4. Insights into the DGA Landscape

4.5 – *Command & Control Priority*

– 23 botnets usa somente o C&C mecanismo de comunicação.

– 5 famílias usam o DGA como canal de comunicação

5. DGA Domain Usage

5.1 – *WHOIS Dataset*

– as análises realizadas foram com base no *dataset DomainTools WHOIS* (9 bilhões de registros).

– 303.165 DGAs relatados para 115.387 domínios.

These WHOIS records contain the following fields

- Date of the WHOIS record
- Date of updates to the WHOIS record
- Dates for domain creation and expiration
- Registrar name
- Registrant name and e-mail address
- Nameservers registered in WHOIS

5. DGA Domain Usage (Family Activity Periods)

Name	T_{first}	T_{last}	$\|S\|$	$\|D_{gen}\|$	$\|D_{uniq}\|$	$\|R_{all}\|$	$\frac{\|R_{all}\|}{\|D_{uniq}\|}$	$\|R_P\|$	$\|R_M\|$	$\|R_S\|$
Kraken	2007-07*	-	1	300	300	-	-	-	-	-
Torpig	2008-01	2011-06-22	2	17,610	17,610	139	0.79%	2	1	57
Szribi	2008-11	2011-06-22	1	4,396	2,949	54	1.83%	0	8	40
Conficker	2008-11*	2009-05-03†	3	129,807,750	125,118,625	-	-	-	-	-
Pushdo TID	2009-07	2012-04-06	1	6,000	6,000	245	4.08%	0	0	0
Pykspa 1	2009-10	2012-09-09	1	32,920	22,764	455	2.00%	12	0	49
Gozi	2010-01	2015-09-18	9	21,890	16,963	305	1.80%	48	4	143
Murofet 1	2010-08	2011-09-08	2	4,063,680	4,063,680	3,172	0.08%	0	50	369
Bamital	2010-11	2013-02-06†	1	197,600	197,600	8,340	4.22%	0	150	30 (7,891)
Nymaim	2011-06	2015-09-16	3	277,112	65,040	656	1.01%	70	17	388
Simda	2011-06	2014-11-06	12	13,000	11,528	379	3.29%	66	9	44
Ramnit	2011-06	2015-02-07	18	18,000	18,000	939	5.22%	0	126	372
Virut	2011-08*	-	1	16,140,000	15,355,008	-	-	-	-	-
Murofet 2	2011-09	2011-12-20	1	262,000	262,000	559	0.21%	0	4	261
Gameover P2P	2011-09	2014-05-28†	1	262,000	262,000	74,755	28.53%	0	23	391 (72,713)
Feodo	2012-02	2012-10-06	3	192	192	110	57.29%	0	1	9
Gootkit	2012-06	2013-11-08	1	2,190	730	198	27.12%	0	0	4
Redyms	2012-12	2014-02-10	1	34	34	11	32.35%	0	2	2
Necurs	2013-01	2015-06-12	6	3,551,232	3,551,232	295	0.01%	10	0	158
CryptoLocker	2013-01	2014-05-30†	1	1,108,000	1,108,000	3,820	0.34%	0	341	240 (2,899)
Suppobox	2013-02	2015-09-20	3	545,169	98,304	11,338	11.53%	8,434	19	792
Banjori	2013-03	2013-09-10	30	434,556	421,390	683	0.16%	0	3	33
Pushdo	2013-03	2015-08-05	4	124,080	124,021	453	0.37%	3	0	54
Pykspa 2	2013-04	2013-10-01	2	775,400	775,342	1,927	0.25%	757	5	101
VolatileCedar	2013-04	2015-03-30	1	170	170	13	7.65%	0	0	7
DirCrypt	2013-07	2014-06-15	14	420	420	86	20.48%	0	13	21
Hesperbot	2013-07	2015-01-07	3	178	178	15	8.43%	0	1	10
Ramdo	2013-10	2014-05-03	3	3,000	3,000	47	1.57%	0	5	23
UrlZone	2013-11	2015-09-20	6	12,006	10,009	127	1.27%	0	24	34
QakBot	2013-12	2015-09-20	1	385,000	385,000	1,088	0.28%	0	61	35
Matsnu	2014-01	2015-09-20	2	3,375	3,346	610	18.23%	244	33	61
Dyre	2014-06	2015-08-19	1	592,000	592,000	850	0.14%	0	1	273
Gameover DGA	2014-07	2014-11-21	2	6,182,000	6,182,000	1,081	0.02%	0	14	549
TinyBanker	2014-08	2015-09-21	90	84,291	81,930	1,733	2.12%	0	272	326
Geodo	2014-10	2014-11-16	2	90,240	90,232	107	0.12%	0	0	39
Tempedreve	2014-10	2015-04-19	1	204	204	20	9.80%	0	0	13
Mewsei	2014-10*	-	1	1,984	1,984	-	-	-	-	-
Fobber	2014-10	2015-07-01	2	2,000	2,000	13	0.65%	0	2	4
Ranbyus	2015-01	2015-08-10	7	105,840	64,400	98	0.15%	0	0	36
Rovnix	2015-01*	-	1	10,000	10,000	1	0.01%	0	0	1
Bedep	2015-02	2015-09-20	4	3,906	3,806	654	17.18%	0	10	201
Corebot	2015-06*	-	2	18,160	18,160	-	-	-	-	-
Shifu	2015-07	2015-09-10	2	1,554	1,554	11	0.71%	0	0	8
Aggregated	-	-	253	165,161,439	159,712,234	115,387	0.63%	9,646	1,199	5,177 (83,503)

Table 3: Overview of DGA usage characteristics. $|S|$: seeds known for this DGA. D: domains generated until 31.12.2015, R: registered domains ($|R_P|$: prior to T_{first}; $|R_M|$: turned into a sinkhole; $|R_S|$: directly registered as sinkhole; in brackets: related to botnet takedowns). Registration percentage based on D_{uniq^*}, thus considering only AGDs where registration data was available.

5. DGA Domain Usage

5.3 – *Domain Registration Status*

- *Pre-registered domains* – 30,25% antes de Tfirst

- *Mitigated domains* – 3,76% domínio transformado em um *sinkhole*.

- *Pure sinkhole domains* – 16,24% registrado por um operador *sinkhole.*

- *Remaining domains* – 49,75% registrados entre Tfirst e Tlast e nenhuma atividade como *sinkhole* (privacidade do WHOIS).

5. DGA Domain Usage (Domain Collisions)

Name	T_{first}	T_{last}	$\|S\|$	$\|D_{gen}\|$	$\|D_{uniq}\|$	$\|R_{all}\|$	$\frac{\|R_{all}\|}{\|D_{uniq}\|}$	$\|R_P\|$	$\|R_M\|$	$\|R_S\|$
Kraken	2007-07*	-	1	300	300	-	-	-	-	-
Torpig	2008-01	2011-06-22	2	17,610	17,610	139	0.79%	2	1	57
Szribi	2008-11	2011-06-22	1	4,396	2,949	54	1.83%	0	8	40
Conficker	2008-11*	2009-05-03†	3	129,807,750	125,118,625	-	-	-	-	-
Pushdo TID	2009-07	2012-04-06	1	6,000	6,000	245	4.08%	0	0	0
Pykspa 1	2009-10	2012-09-09	1	32,920	22,764	455	2.00%	12	0	49
Gozi	2010-01	2015-09-18	9	21,890	16,963	305	1.80%	48	4	143
Murofet 1	2010-08	2011-09-08	2	4,063,680	4,063,680	3,172	0.08%	0	50	369
Bamital	2010-11	2013-02-06†	1	197,600	197,600	8,340	4.22%	0	150	30 (7,891)
Nymaim	2011-06	2015-09-16	3	277,112	65,040	656	1.01%	70	17	388
Simda	2011-06	2014-11-06	12	13,000	11,528	379	3.29%	66	9	44
Ramnit	2011-06	2015-02-07	18	18,000	18,000	939	5.22%	0	126	372
Virut	2011-08*	-	1	16,140,000	15,355,008	-	-	-	-	-
Murofet 2	2011-09	2011-12-20	1	262,000	262,000	559	0.21%	0	4	261
Gameover P2P	2011-09	2014-05-28†	1	262,000	262,000	74,755	28.53%	0	23	391 (72,713)
Feodo	2012-02	2012-10-06	3	192	192	110	57.29%	0	1	9
Gootkit	2012-06	2013-11-08	1	2,190	730	198	27.12%	0	0	4
Redyms	2012-12	2014-02-10	1	34	34	11	32.35%	0	2	2
Necurs	2013-01	2015-06-12	6	3,551,232	3,551,232	295	0.01%	10	0	158
CryptoLocker	2013-01	2014-05-30†	1	1,108,000	1,108,000	3,820	0.34%	0	341	240 (2,899)
Suppobox	2013-02	2015-09-20	3	545,169	98,304	11,338	11.53%	8,434	19	792
Banjori	2013-03	2013-09-10	30	434,556	421,390	683	0.16%	0	3	33
Pushdo	2013-03	2015-08-05	4	124,080	124,021	453	0.37%	3	0	54
Pykspa 2	2013-04	2013-10-01	2	775,400	775,342	1,927	0.25%	757	5	101
VolatileCedar	2013-04	2015-03-30	1	170	170	13	7.65%	0	0	7
DirCrypt	2013-07	2014-06-15	14	420	420	86	20.48%	0	13	21
Hesperbot	2013-07	2015-01-07	3	178	178	15	8.43%	0	1	10
Ramdo	2013-10	2014-05-03	3	3,000	3,000	47	1.57%	0	5	23
UrlZone	2013-11	2015-09-20	6	12,006	10,009	127	1.27%	0	24	34
QakBot	2013-12	2015-09-20	1	385,000	385,000	1,088	0.28%	0	61	35
Matsnu	2014-01	2015-09-20	2	3,375	3,346	610	18.23%	244	33	61
Dyre	2014-06	2015-08-19	1	592,000	592,000	850	0.14%	0	1	273
Gameover DGA	2014-07	2014-11-21	2	6,182,000	6,182,000	1,081	0.02%	0	14	549
TinyBanker	2014-08	2015-09-21	90	84,291	81,930	1,733	2.12%	0	272	326
Geodo	2014-10	2014-11-16	2	90,240	90,232	107	0.12%	0	0	39
Tempedreve	2014-10	2015-04-19	1	204	204	20	9.80%	0	0	13
Mewsei	2014-10*	-	1	1,984	1,984	-	-	-	-	-
Fobber	2014-10	2015-07-01	2	2,000	2,000	13	0.65%	0	2	4
Ranbyus	2015-01	2015-08-10	7	105,840	64,400	98	0.15%	0	0	36
Rovnix	2015-01*	-	1	10,000	10,000	1	0.01%	0	0	1
Bedep	2015-02	2015-09-20	4	3,906	3,806	654	17.18%	0	10	201
Corebot	2015-06*	-	2	18,160	18,160	-	-	-	-	-
Shifu	2015-07	2015-09-10	2	1,554	1,554	11	0.71%	0	0	8
Aggregated	-	-	253	165,161,439	159,712,234	115,387	0.63%	9,646	1,199	5,177 (83,503)

Table 3: Overview of DGA usage characteristics. $|S|$: seeds known for this DGA. D: domains generated until 31.12.2015, R: registered domains ($|R_P|$: prior to T_{first}; $|R_M|$: turned into a sinkhole; $|R_S|$: directly registered as sinkhole; in brackets: related to botnet takedowns). Registration percentage based on D_{uniq^*}, thus considering only AGDs where registration data was available.

5. DGA Domain Usage (Domain Registration Lookahead)

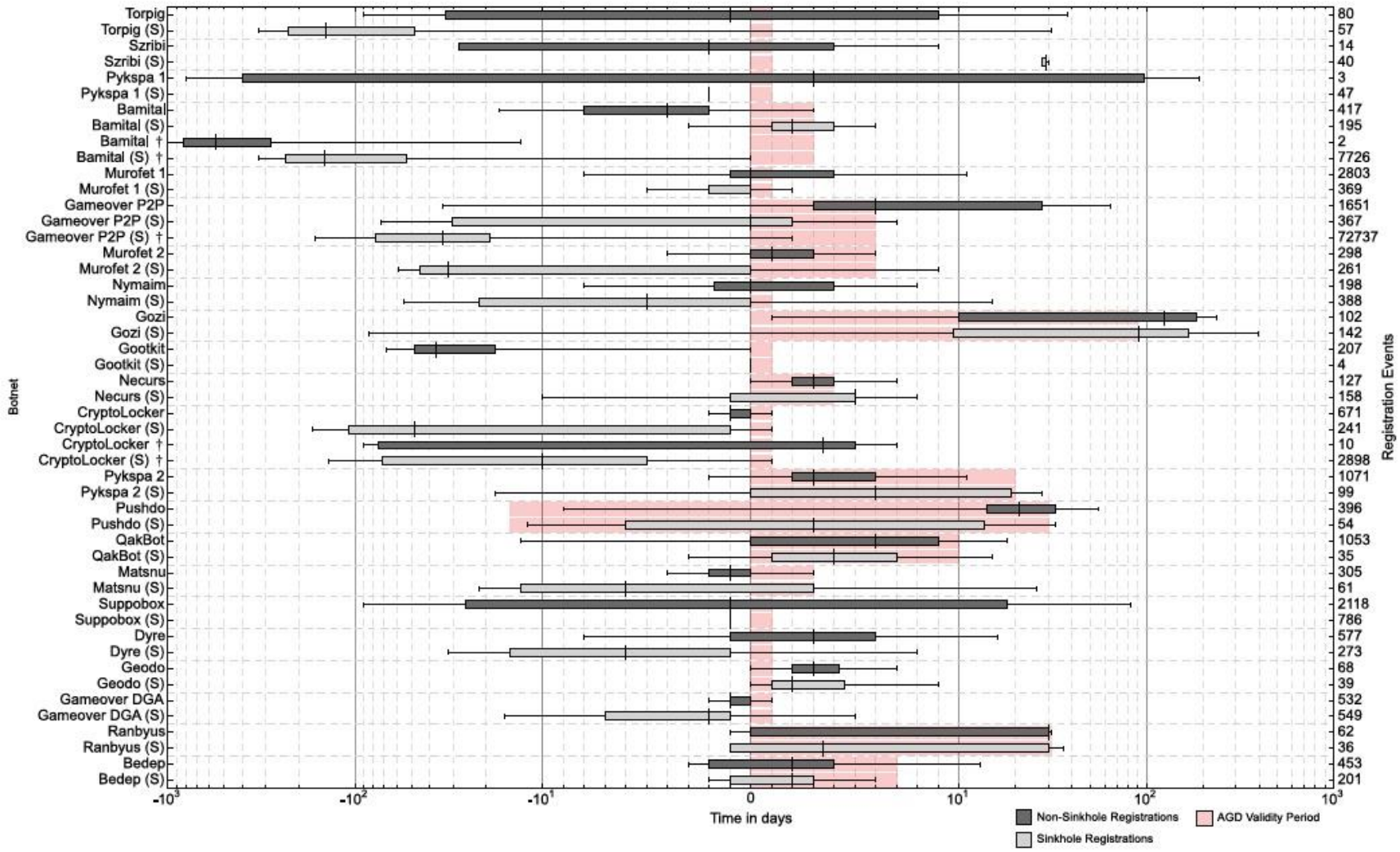

Figure 2: Lookahead of domain registrations in time-dependent DGAs, divided into identifiable sinkholes and remaining domains. The data in this boxplot is applied relative to the start the respective AGD's validity period (shown in light red for better orientation). Bamital, GameoverP2P, CryptoLocker data is further divided into pre and post takedown (indicated by †).

5. DGA Domain Usage

5.6 – *Mitigation Response Time*

– mudança de um operador normal para um *sinkhole*.

Name	Seeds	Mitigations	Validity Period		First Reponse (in days)				Further Reponses (in days)			
			within	after	m_{min}	$\tilde{m}$	$\bar{m}$	m_{max}	m_{min}	$\tilde{m}$	$\bar{m}$	m_{max}
Murofet 1	1	50	37	13	25	25	25.00	25	0	0	1.77	25
Bamital	1	148	16	132	6	6	6.00	6	3	49	47.37	92
Nymaim	2	16	4	12	1	16	16.00	31	0	5	31.46	212
Ramnit	16	126	-	-	2	19	34.60	153	0	35	54.28	363
CryptoLocker	1	216	25	191	9	9	9.00	9	0	8	14.69	130
Suppobox	2	19	13	6	3	117	117.50	232	0	0	25.12	194
DirCrypt	7	13	-	-	0	3	7.00	26	0	5	10.50	41
UrlZone	4	24	-	-	0	4	6.75	19	0	114	112.40	251
QakBot	1	33	15	18	0	0	0.00	0	0	21	28.41	66
Matsnu	2	33	11	22	2	2	2.50	3	2	7	9.16	72
Gameover DGA	2	14	9	5	0	1	1.50	3	0	1	54.17	161
TinyBanker	51	272	-	-	0	3	6.39	61	0	2	5.66	60
Bedep	2	10	4	6	2	5	5.00	8	1	12	13.12	28

Table 4: Mitigation Response Timings for selected DGAs. For time-dependent DGAs, *Validity Period* describes the identified mitigations for active and outdated AGDs. *First Response* is the time until the first mitigation occurred, with values for minimum, median, average, and maximum, aggregated over seeds. *Further Responses* describes the same measures for all following events.

5. DGA Domain Usage

5.7 – *DGAs and Domain Parking*

– *parking*: após o período de registro terminar, o domínio expirado é transferido ou escolhido por um revendedor de domínios.

– Banjori 620 (90,78%)

– QakBot 959 (54,69%)

– Pykspa 2 883 (45,82%)

– 2.917 (87,52%) domínios registrados com o mesmo serviço de *parking domain*.

5. DGA Domain Usage

5.8 – *Discussion: Countering DGAs*

– 3 word-list based (domínios com significados)

– DGAs se tornaram muito importantes para os autores de *malwares* (25 em 2 anos).

– assimetria: um *attacker* precisa de um único domínio válido para garantir o controle de uma rede, um *defender* precisa bloquear todos os domínios em potencial.

– financeiramente custoso para realizar um *takedown.*

5. DGA Domain Usage

5.8 – Discussion: Countering DGAs

– domínios registrados com *takedown* são utilizados como *sinkholes.*

– 3.302 colisões entre 5 DGAs entre 159.712.234 domínios únicos gerados.

6. Related Work

Faz o levantamento do estado da arte sobre o assunto estudado.

– Barabosch definiu a taxonomia dos tipos de DGAs baseada em 2 características: *time-dependence* e *causality*.

7. Conclusion

– a maior descoberta é que que esse dataset de domínios pode ser usado tanto para bloqueio preventivo de tentativas de acessos C&C, quanto melhorar as famílias de *malwares* e companhias com basicamente não falso positivo.

– caracterizou-se o comportamento de registro de *botmasters* e de *sinkholers* e examinou a efetividade de *domain mitigation*.

– como contribuição: https://dgarchive.caad.fkie.fraunhofer.de.

Referência

- A Comprehensive Measurement Study of Domain Generating Malware
- http://www.alexa.com/topsites
- http://whois.domaintools.com/