Detecting Fake Medical Web Sites Using Recursive Trust Labeling

AHMED ABBASI, University of Virginia
FATEMEH "MARIAM" ZAHEDI, University of Wisconsin-Milwaukee
SIDDHARTH KAZA, Towson University

Fake medical Web sites have become increasingly prevalent. Consequently, much of the health-related information and advice available online is inaccurate and/or misleading. Scores of medical institution Web sites are for organizations that do not exist and more than 90% of online pharmacy Web sites are fraudulent. In addition to monetary losses exacted on unsuspecting users, these fake medical Web sites have severe public safety ramifications. According to a World Health Organization report, approximately half the drugs sold on the Web are counterfeit, resulting in thousands of deaths. In this study, we propose an adaptive learning algorithm called recursive trust labeling (RTL). RTL uses underlying content and graph-based classifiers, coupled with a recursive labeling mechanism, for enhanced detection of fake medical Web sites. The proposed method was evaluated on a test bed encompassing nearly 100 million links between 930,000 Web sites, including 1,000 known legitimate and fake medical sites. The experimental results revealed that RTL was able to significantly improve fake medical Web site detection performance over 19 comparison content and graph-based methods, various meta-learning techniques, and existing adaptive learning approaches, with an overall accuracy of over 94%. Moreover, RTL was able to attain high performance levels even when the training dataset composed of as little as 30 Web sites. With the increased popularity of eHealth and Health 2.0, the results have important implications for online trust, security, and public safety.

Categories and Subject Descriptors: I.2.6 [**Artificial Intelligence**]: Learning—*Knowledge acquisition*; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Information filtering; Selection process*

General Terms: Design, Algorithms, Experimentation, Performance, Security

Additional Key Words and Phrases: Fake Web sites, Web spam, machine learning, Web mining, Health 2.0, medical fraud

ACM Reference Format:
Abbasi, A., Zahedi, F. M., and Kaza, S. 2012. Detecting fake medical Web sites using recursive trust labeling. ACM Trans. Inf. Syst. 30, 4, Article 22 (November 2012), 36 pages.
DOI = 10.1145/2382438.2382441 http://doi.acm.org/10.1145/2382438.2382441

1. INTRODUCTION

Fake Web sites are misrepresentative sites posing as legitimate online sources of information, goods, and/or services [Abbasi and Chen 2009a]. The three major types of fake Web sites (spoof, concocted, and Web spam) collectively generate billions of dollars in fraudulent revenue [Dinev 2006]. Spoof Web sites engage in identity theft by mimicking legitimate Web sites and targeting those Web sites' customers through

This work is supported by the National Science Foundation, under grant CNS-1049497.
Authors' addresses: A. Abbasi, Information Technology, University of Virginia; email: abbasi@comm.virginia.edu; F. M. Zahedi, Information Technology and Management, University of Wisconsin-Milwaukee; S. Kaza, Department of Computer and Information Sciences, Towson University.

DOI 10.1145/2382438.2382441 http://doi.acm.org/10.1145/2382438.2382441

phishing emails [Dinev 2006; Fu et al. 2006]. Concocted Web sites attempt to appear as unique, legitimate commercial entities in order to engage in failure-to-ship fraud [Abbasi and Chen 2009a; Chua and Wareham 2004; Chua et al. 2007]. Web spam sites engage in black-hat search engine optimization (using content and/or link spamming techniques) in order to bolster their search rank, often for commercial incentives [Gyongyi and Garcia-Molina 2005].

These three types of fake Web sites span numerous domains [Abbasi et al. 2010; Grazioli and Jarvenpaa 2000]. One emerging domain is fake medical Web sites; which include fraudulent online pharmacies, health information providers, and medical institution Web sites. Fake medical Web sites have important implications for online trust and security, particularly since people are increasingly turning to the Web as a source for health-related information and products [Song and Zahedi 2007; White and Horvitz 2009]. While the primary cost exacted on Internet users by fake Web sites can be quantified in monetary terms, in the case of fake medical Web sites, these monetary costs are coupled with dire social ramifications. For instance, online pharmacies that fail to ship essential medication (or sell counterfeit drugs) and Web sites intentionally providing inaccurate or fictitious medical information have major implications pertaining to Internet users' health and wellness [Hesse et al. 2010]. According to studies conducted by the U.S. Food and Drug Administration and the World Health Organization, thousands of deaths have been attributed to fake medical Web sites, while the number of people visiting such sites continues to increase dramatically [Easton 2007; Krebs 2005].

Prior fake Web site detection research has focused on the application of content or graph-based detection methods to concocted or spoof e-commerce Web sites targeting end users [Abbasi and Chen 2009a; Liu et al. 2006] or Web spam sites targeting search engines [Becchetti et al. 2008; Wang et al. 2008]. The lack of prior work on fake medical Web sites (despite their increased prevalence), coupled with the public's growing reliance on them, point to an urgent need to evaluate the efficacy of detection methods for such Web sites [Luo 2008; Zahedi and Song 2008]. Considering that in the case of fake medical Web sites, improvements in detection have important social implications stemming from Internet user's health and wellness, a comprehensive examination of the performance of state-of-the-art content and graph-based techniques is well motivated.

In this study, we propose an adaptive learning algorithm called recursive trust labeling (RTL). RTL uses underlying content and graph-based classifiers designed to exploit the unique characteristics of fake medical Web sites, coupled with a recursive labeling mechanism. Given the complexities associated with medical content, the content classifier incorporates several novel components, including a medical thesaurus. Similarly, RTL's graph classifier employs characteristics closely aligned with the unique linkage tendencies exhibited by medical Web sites. The recursive labeling mechanism effectively leverages the information provided by these two classifiers towards enhanced detection of fake medical Web sites. The proposed algorithm was compared against numerous existing content and graph-based methods. Experimental results revealed that RTL yielded more accurate results than comparison classifiers and also outperformed meta-learning strategies such as stacking. RTL was more effective than existing adaptive learning methods even when the quantity of training data was very small and imbalanced; a situation that is often encountered when detecting fake Web sites. RTL's improved performance was attributable to its ability to better leverage content and linkage-related characteristics of online medical content (through the content and graph classifiers), as well as the effectiveness of the recursive labeling mechanism. Given the hefty social cost exacted by fake medical Web sites, the results of this study have important implications.

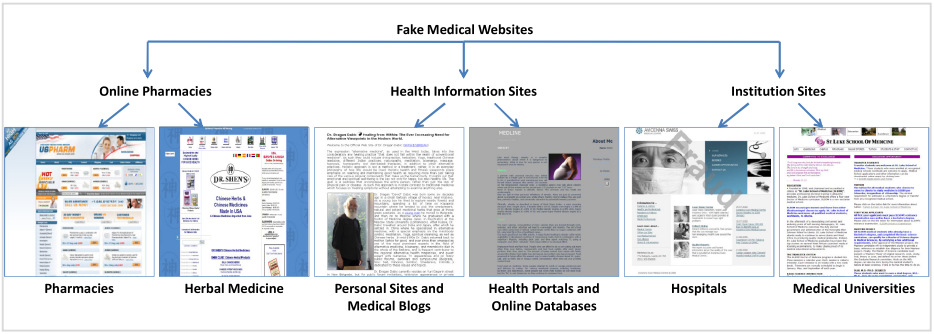

Fig. 1. Categories of fake medical Web sites.

2. FAKE MEDICAL WEB SITES

There has been a growing emphasis on the development of information technologies capable of providing users with easy access to the copious amounts of medical content available on the Web [Chen et al. 2003; Gaudinat et al. 2007; Luo 2008]. These technologies, which are critical to the continued success of eHealth and Health 2.0, are highly susceptible to the prevalence of fake medical Web sites, resulting in major information quality degradation and public safety concerns [Eysenbach 2008; Hesse et al. 2010; Hughes et al. 2008; Price and Hersh 1999]. Figure 1 shows examples of different categories of fake medical Web sites, with example screenshots of various fake sites. Common categories include fake online pharmacies, fake health information providers (e.g., infomediaries and medical blogs), and phony institution Web sites (e.g., medical universities, hospitals, clinics, etc.).

Fake online pharmacy Web sites sell fake drugs and/or fail to ship the agreed-upon goods altogether [Armin 2010; Easton 2007; Greenberg 2008]. Due to the growth in the usage of Internet pharmacies [Boggan 2009; Wilford et al. 2006], fake pharmacy Web sites have become highly pervasive. According to a Food and Drug Administration study, more than 90% of the 12,000 Internet pharmacies examined were fraudulent [Krebs 2005]. As a result, the number of users visiting fake pharmacy sites has tripled since 2008; on average, fake online pharmacies each get 100,000 hits per year [Greenberg 2008]. Consequently, according to the World Health Organization, fake online pharmacies are responsible for thousands of deaths [Boggan 2009; Easton 2007]. In one instance, investigators in the UK found fake pills laced with rat poison, cement, floor polish, chalk, rice flour, and lead paint [Boggan 2009].

Fake health information Web sites provide fictitious, false, misleading, or biased medical information about nutrition, exercise, diet, treatment options, surgery, drug side effects, miracle remedies, hospital rankings, etc. [Aphinyanaphongs and Aliferis 2007; Wang and Richard 2007]. These Web sites, which are often linked to fake online pharmacies and health/medical equipment sellers, are commonly used as an advertising and propaganda dissemination mechanism. Fake health information Web sites are especially problematic since recent research has noted that while 61% of Americans search for healthcare information online, 75% do not check the validity of the information sources [Pew Internet and American Life Project 2009; White and Horvitz 2009].

Fake medical institution Web sites include those attempting to appear as legitimate hospitals or medical universities. These Web sites are often used for medical identity theft or to defraud individuals suffering from specific ailments [Parloff 2010]. Table I presents a summary of the three major categories of fake medical Web sites, including examples of potential ramifications. The grave social implications associated with fake

Table I. Summary of Fake Medical Web Site Categories

Category	Problem Description	Fraud Examples
Online Pharmacies	Engage in failure-to-ship fraud and/or the sale of counterfeit drugs. According to a 2005 FDA study, nearly 11,000 of the 12,000 online pharmacies examined were fake.	A 58-year-old Canadian woman died after taking fake sleeping pills laced with strontium, arsenic, aluminum, and uranium.
		An American woman died after taking medicine laced with aluminum.
		A 22-year-old British girl died after taking fake medication for anxiety, stomachache, and insomnia.
Health Information Providers	Provide inaccurate, misleading, or fictitious health related information. According to a 2006 Pew Research Center report, 75% of Americans do not check the validity of online healthcare information sources.	The developer of the alternative medicine Web site letstalkhealth.com was arrested in 2009 on various felony charges, including falsely representing a cure for cancer.
Medical Institutions	Engage in medical identity theft (including insurance and medicare information) and/or monetary fraud.	A group of 14 fraudsters in China developed 7 fake military hospital Web sites used to defraud nearly 10,000 people.
		Fake US veterans' hospital Web sites were used to deceive former soldiers suffering from asbestos exposure related illnesses.

Sources: [An 2010; Armin 2010; Boggan 2009; Parloff 2010]

medical Web sites make their accurate detection an issue of paramount importance [Aldhous 2005].

Online communities and regulatory agencies that maintain databases of known fake Web sites based on manual observations are not capable of keeping up with the plethora of fakes emerging on a daily basis [Abbasi et al. 2010]. For instance, the National Association of Boards of Pharmacies' list of fraudulent online pharmacies only covers a small percentage of the actual number in existence. Consequently, prior research has focused on automated learning-based approaches for detecting fake Web sites. A review of prior work on fake Web site detection is presented in the ensuing section. We present our approach for fake medical Web site detection in Section 4.

3. FAKE WEB SITE DETECTION

Two categories of automated learning-based techniques have been applied to fake Web site detection: content and graph-based methods. Content-based methods utilize machine learning algorithms coupled with content-based features called fraud cues [Abbasi and Chen 2009b]. Graph-based methods use properties of the Web site hyperlink graphs to deduce whether a particular Web site is legitimate or fake [Gyongyi and Garcia-Molina 2005; Zhang et al. 2006]. These two categories are reviewed, along with prior fusion strategies for combining content and graph information, and adaptive learning methods.

3.1. Content-Based Methods

Fraudsters frequently expedite the development of fake Web sites by using automatic content generation techniques to mass produce Web pages [Urvoy et al. 2008]. This often causes fake Web sites to appear templatic: that is, there are potential content-based design similarities between new and existing fake Web sites [Fetterly et al. 2004]. These similarities (called "fraud cues"), can be exploited using machine

Table II. Selected Prior Studies on Content-Based Methods for Fake Web Site Detection

Study	Site Type	Features	Techniques	Test Bed and Evaluation Results
Chou et al. 2004	Spoof	HTML tags, URL text, image hashes, link information	Test based scoring mechanism (called TSS)	719 legit and fake Web sites; 67.8% accuracy
Drost and Scheffer 2005	Web Spam	Body text tokens, redirections, characters in URLs, number of in/out links and their content	SVM (linear, polynomial, RBF kernels)	Web pages (quantity used unclear); over 95% accuracy
Fu et al. 2006	Spoof	Web page snapshot	Earth Mover's Distance algorithm	10,272 legit and 9 fake pages;
Liu et al. 2006	Spoof	Body text, layout, and images	Visual similarity assessment algorithm	328 Web pages (320 legit, 8 spoof);
Ntoulas et al. 2006	Web Spam	Body text lexical measures and n-grams, anchor text	C4.5, Neural Network, SVM	Over 17,000 Web pages; 95.4% accuracy
Shen et al. 2006	Web Spam	Temporal link features such as in link growth/death rate	SVM (linear kernel)	113,756 Web pages;
Urvoy et al. 2006	Web Spam	HTML style markers	Jaccard based similarity algorithm (called HSS)	5 million Web pages;
Wu and Davison 2006	Web Spam	Body text tokens, HTML tags, in/out links, relative/absolute links	SVM (linear kernel)	1,285 Web pages; 93% precision, 85% recall
Urvoy et al. 2008	Web Spam	HTML style markers	Jaccard based similarity algorithm (called HSS)	5,400 legit and spam hosts; f-measure between 55%-60%
Abbasi and Chen 2009b	Concocted	Body, HTML, and URL text; images, number of in/out links	C4.5, Naïve Bayes, SVM (linear kernel), Winnow	350 legit and fake Web sites; 96.7% accuracy
Martinez-Romo and Araujo 2009	Web Spam	Body text tokens, URL and anchor text tokens, and language models	Kullback-Leibler divergence	3083 legit and spam hosts; 81% f-measure
Abbasi et al. 2010	Concocted, Spoof	Body text lexical measures and n-grams, HTML and URL and anchor text n-grams, image pixels, number of in/out links	Bayesian Network, C4.5, Logit Regression, Naïve Bayes, Neural Network, SVM (linear composite, linear, polynomial, RBF kernels),	900 legit, concocted, and spoof e-commerce Web sites; 92.56% accuracy
Le et al. 2011	Spoof	URL tokens and lexical and syntactic measures, domain registration information	SVM (linear kernel), online learning algorithms	14,238 legit and spoof URLs; 96.86% accuracy

learning algorithms [Abbasi and Chen 2009b; Drost and Scheffer 2005]. Table II presents a summary of select studies that used content-based methods for detection of Web spam, spoof, and concocted Web sites. While the table is not an exhaustive list of prior content-based studies, it provides important insights into the feature

representations and classification methods utilized. For instance, prior studies have often employed n-grams derived from the Web pages' body text, URLs, and source code as the feature representations [Abbasi and Chen 2009b; Chou et al. 2004; Drost and Scheffer 2005; Le et al. 2011; Ntoulas et al. 2006; Wu et al. 2006].

Based on Table II, it is apparent that various machine learning techniques have been used in previous fake Web site classification research, including Support Vector Machines (SVM), Neural Networks, Bayesian Networks, Naïve Bayes, C4.5, and Logistic Regression. SVM has been particularly effective in numerous prior Web spam categorization studies. Drost and Scheffer [2005] attained over 95% accuracy using linear and radial basis function (RBF) kernel SVMs to differentiate ham pages from spam. Shen et al. [2006] trained a linear SVM using temporal features for Web spam categorization. SVM has also worked well on detection of concocted escrow Web sites [Abbasi and Chen 2009b]. Moreover, it has attained good results on related work pertaining to detection of blog spam (i.e., splogs) [Kolari et al. 2006]. For instance, Lin et al. [2007] achieved over 95% accuracy for Weblog and splog categorization using body, URL, and anchor text coupled with temporal features and an RBF SVM kernel. SVM has also been used for predicting the reliability of medical pages [Sondhi et al. 2012].

Other relevant classification algorithms include the C4.5 decision tree [Abbasi and Chen 2009b; Ntoulas et al. 2006]. C4.5 uses the information gain heuristic to select attributes which provide the highest entropy reduction on the training data [Quinlan 1986]. These features are used to build a decision tree model. Based on Bayes' Theorem [Bayes 1958], Bayesian Networks and Naïve Bayes have both been used for fake Web site detection. Naïve Bayes is a fairly simple probabilistic classification algorithm that uses strong independence assumptions regarding various features. Since these assumptions enable it to efficiently build models, it has been utilized for Web spam, concocted, and spoof Web site detection [Abbasi et al. 2010; Salvetti and Nicolov 2006]. Neural Networks and Logistic Regression have also been applied to Web spam, concocted, and spoof Web sites [Abbasi et al. 2010; Ntoulas et al. 2006]. In recent comparisons on concocted and spoof Web sites, Bayesian Network, Logistic Regression, and C4.5 all performed well [Abbasi et al. 2010]. Other studies have used body text, URL, HTML, and image features with similarity measures such as Jaccard, Kullback-Leibler, and Earth Mover's Distance for spoof and Web spam detection [Liu et al. 2006; Martinez-Romo and Araujo 2009; Urvoy et al. 2006, 2008].

3.2. Graph-Based Methods

Fake medical Web site developers routinely spend millions of dollars on black-hat search engine optimization in order to make their Web sites more visible [Greenberg 2008]. This is commonly accomplished through the use of link farms; numerous Web sites pointing to a set of designated sites, in order to artificially inflate their perceived importance [Gyongyi and Garcia-Molina 2005; Wang et al. 2008; Wu and Davison 2005]. Graph-based techniques designed to combat link farms rely on the assumption that good pages link to good pages, while bad pages are likely to link to other bad ones. This notion of using linkage information to assess the quality of a Web page is embodied by the Page Rank algorithm [Page et al. 1998]. Consequently, most graph-based fake Web site detection algorithms follow the Page-Rank intuition by iteratively traversing links in the Web graph while constantly updating each node's score. During the traversal process, these methods also propagate trust and/or distrust from a set of seed URLs that are known to be good or bad. Several graph-based techniques have been proposed for detecting fake Web sites. Tables III and IV present a summary of the methods incorporated in this study. They can be categorized according to their

Table III. Single-Class Propagation Methods Employed in this Study

Algorithm	Seed Pages	Propagation
TrustRank [Gyongyi et al. 2004]	E(A) = 1, for known good pages E(A) = 0, for all other pages	$TR(A) = \frac{E(A)(1-d)}{n} + d\sum_{i=1}^{n}\frac{TR(I_i)}{C(I_i)}$ where: E(A) is the initial TrustRank score of A $TR(I_i)$ is the TrustRank score of I_i, an inlink of A $C(I_i)$ is the is the number of outlinks for I_i d is a tunable parameter between 0 and 1
BadRank [Wu and Davison 2005]	E(A) = 1, for known bad pages E(A) = 0, for all other pages	$BR(A) = E(A)(1-d) + d\sum_{i=1}^{n}\frac{BR(T_i)}{C(T_i)}$ where: E(A) is the initial BadRank score of A $BR(T_i)$ is the BadRank score of T_i, an inlink of A $C(T_i)$ is the is the number of inlinks for T_i d is a tunable parameter between 0 and 1
Parent Penalty [Wu and Davison 2005]	S(A) = 1, if number of common sites in A's inlinks and outlinks exceeds a threshold t S(A) = 0, for all other pages	$S(A) = 1$, if $\sum_{i=1}^{n} S(T_i) \geq p$ S(A) = 0, otherwise where: $S(T_i)$ is the score of T_i, an outlink of A, p is a parameter
Anti TrustRank [Krishnan and Raj 2006]	E(A) = 1, for known bad pages E(A) = 0, for all other pages	$AR(A) = \frac{E(A)(1-d)}{n} + d\sum_{i=1}^{n}\frac{AR(I_i)}{C(I_i)}$ where: E(A) is the initial AntiTrustRank score of A $AR(I_i)$ is the AntiTrustRank score of I_i, an inlink of A $C(I_i)$ is the number of outlinks for I_i d is between 0 and 1
Mass Estimation [Gyongyi et al. 2006]	Uses known good pages to compute TR(A), the TrustRank score of A	$SM(A) = \frac{PR(A)-TR(A)}{PR(A)}$ where: SM(A) is the spam mass of A, PR(A) is the PageRank
Cautious Surfer [Nie et al. 2007b]	Uses known good pages to compute TR(A), the TrustRank score of A	$T(A) = 1 - \text{rank}\left(\frac{TR(A)}{n}\right)$ $CS(A) = \sum_{i=1}^{k}\frac{CS(I_i)T(I_i)}{\sum_{j=1}^{p}T(O_j)} + \sum_{m=1}^{n}\frac{1-CS(1_m)T(1_m)}{T(1_m)}$ where: $CS(I_i)$ is the cautious surfer score of I_i, an inlink of A O_j is one of the p outlinks of I_i, 1_m is one of the n nodes

propagation mechanisms, which are based on the assumptions they make regarding how real and fake Web pages are connected to one another.

Single class propagation algorithms begin with a seed set of known good or bad pages (Table III). These pages' "goodness" or "badness" is then propagated through the graph via their inlinks and/or outlinks. Examples of single class propagation algorithms include BadRank, TrustRank, AntiTrustRank, Mass Estimation, and ParentPenalty [Gyongyi and Garcia-Molina 2005; Gyongyi et al. 2004; Krishnan and Raj 2006; Wu and Davison 2005]. BadRank, AntiTrustRank, and ParentPenalty all propagate distrust through known bad pages [Krishnan and Raj 2006; Wu and Davison 2005]. In contrast, TrustRank, Mass Estimation, and Cautious Surfer all relay trust through good pages [Gyongyi and Garcia-Molina 2005; Gyongyi et al. 2004; Nie et al. 2007b].

Table IV. Unsupervised and Dual-Class Propagation Methods Employed in This Study

Algorithm	Seed Pages	Propagation
PageRank [Page et al. 1998]	Not applicable	$PR(A) = \frac{(1-d)}{n} + d\sum_{i=1}^{n}\frac{PR(I_i)}{C(I_i)}$ where: $PR(I_i)$ is the PageRank score of I_i, an inlink of A $C(I_i)$ is the number of outlinks for I_i d is a tunable parameter between 0 and 1
QoL [Zhang et al. 2006]	EL(A) = 1, if A's good outlinks > bad outlinks, and bad outlinks > k EL(A) = −1, for all other pages	$QoL(A) = EL(A)(1-d) + d\left(\sum_{i=1}^{n}\left(\beta\frac{QoC(T_i)}{C(T_i)} + (1-\beta)\frac{QoL(T_i)}{C(T_i)}\right)\right)$ where: EL(A) is the initial score of A $QoC(T_i)$ is the QoC score of T_i, an outlink of A $QoL(T_i)$ is the QoL score of T_i $C(T_i)$ is the number of inlinks of T_i β and d are tunable parameters between 0 and 1
QoC [Zhang et al. 2006]	EC(A) = 1, for known good pages EC(A) = 0, for all other pages	$QoC(A) = EC(A)(1-d) + d\left(\sum_{i=1}^{n}\left(\alpha\frac{QoC(I_i)}{C(I_i)} + (1-\alpha)\frac{QoL(I_i)}{C(I_i)}\right)\right)$ where: EC(A) is the initial score of A $QoC(I_i)$ is the QoC score of I_i, an outlink of A $QoL(I_i)$ is the QoL score of I_i $C(I_i)$ is the number of inlinks of I_i α and d are tunable parameters between 0 and 1
Trust Distrust [Nie et al. 2007a; Wu et al. 2006]	EC(A) = 1, for known good pages F(A) = 1, for known bad pages E(A), F(A) = 0, for all other pages	$TD(A) = \alpha T(A) - \beta D(A)$ where: $T(A) = \frac{E(A)(1-d)}{n} + d\sum_{i=1}^{n}\frac{T(I_d)}{\log(1+C(I_i))}$ $D(A) = \frac{F(A)(1-d)}{n} + d\sum_{i=1}^{n}\frac{D(O_i)}{\log(1+G(O_i))}$ E(A) and F(A) are the initial trust/distrust scores of A $T(I_i)$ is the trust score of I_i, an inlink of A $D(O_i)$ is the distrust score of O_i, an outlink of A $C(I_i)$ is the number of outlinks for I_i $G(O_i)$ is the number of inlinks for O_i α, β and d are tunable parameters between 0 and 1

Dual class propagation algorithms, such as QoC, QoL, and TrustDistrust all utilize good and bad seed pages and also propagate scores through both inlinks and outlinks [Nie et al. 2007b; Zhang et al. 2006]. A third class of algorithms (e.g., PageRank) do not use any seed pages; they are purely unsupervised techniques that were not designed to detect bad pages [Page et al. 1998]. Nevertheless, PageRank is effective at detecting fake Web pages that do not utilize link farms, and is therefore often used as a baseline [Zhang et al. 2006]. For completeness, the PageRank formulation is included in Table IV prior to the dual-class algorithms.

3.3. Fusion Strategies for Combining Content and Graph Information

Combining content and structural information has yielded good results for various Web mining tasks [Adomavicius and Tuzhilin 2005; Fang et al. 2007]. It can also improve fake Web site detection performance, as shown in prior studies [Abernathy et al. 2008; Araujo and Martinez-Romo 2010]. One commonly used approach is to combine content and link-based features into a single feature vector, which is then input into a classifier [Wu et al. 2006]. A caveat is that representing linkage information in feature vector form can inhibit the representational richness of the link-based information captured. Nevertheless, combining content and link features in such a manner

has resulted in improved fake Web site detection capabilities as compared to using either category of information in isolation [Drost and Scheffer 2005]. For example, certain fake online escrow Web sites (a major category of concocted sites) are easier to detect using content-based features, while others are more susceptible to link-based attributes [Abbasi and Chen 2009b].

Metalearning uses the expertise acquired through underlying machine learning or data mining processes to increase the quality of results obtained in subsequent applications [Brazdil et al. 2008]. Various meta-learning strategies have been utilized for fake Web site detection. Stack and ensemble classifiers have improved results of underlying content-based classifiers, both for concocted and spoof site detection [Abbasi and Chen 2009a, 2009b]. Stack classifiers have also improved performance when used with features derived from underlying graph-based algorithms such as PageRank and TrustRank [Becchetti et al. 2008]. Castillo et al. [2007] used an ensemble of classifiers, each trained with content and link-based features, to improve Web spam detection performance over results attained using a single C4.5 decision tree.

3.4. Using Adaptive Learning to Overcome Data Limitations

Recently, adaptive learning has shown promising results. Reclassifying/relabeling test data instances as new information becomes available (i.e., using other instances' predictions as features) has worked well on Web spam graphs [Gan and Suel 2007]. Similarly, Tian et al. [2007] used semi-supervised learning to generate new link features, which were then added to the original feature set and used to reclassify the testing data. Castillo et al. [2007] used a stacked graphical learning method that improved performance for an underlying classifier when run for two iterations. Due to the difficulties associated with manually identifying and gathering fake Web sites, such adaptive learning approaches are particularly important since the quantity of known good/bad pages is often limited relative to the overall size of the test bed. Given the size of the Web, in many instances, the potential training data constitutes less than 1% of the nodes in the graph [Wu and Chellapilla 2007]. The situation is further exacerbated by the difficulty associated with finding and collecting known fake Web sites that are still alive: many are taken down within a few days after appearing in online fraud prevention communities' databases [Abbasi and Chen 2009a; Zhang et al. 2006]. Therefore, relabeling nodes and retraining models as new information becomes available during the learning process could be highly beneficial [Le et al. 2011].

4. RECURSIVE TRUST LABELING

Leveraging the insights gained from prior studies, we propose an adaptive learning algorithm called Recursive Trust Labeling (RTL), which uses underlying graph and content-based classifiers that exploit the unique characteristics of fake medical Web sites. The recursive labeling mechanism recursively expands the training dataset by selecting additional test instances during each iteration. Instances that have the strongest prediction scores (based on the content and graph classifiers) are selected and added with class labels that are based on the underlying classifiers' predictions. The RTL algorithm is designed to exploit the complementary information utilized by content and graph-based classifiers in a dynamic manner; classifications are revised and improved as new information becomes available.

When using semisupervised learning, a critical problem arises when misclassified instances are added to the training data [Tian et al. 2007]. This is a major concern since classification models can incorporate incorrect rules and assumptions, resulting in amplified error rates [Gan and Suel 2007]. To avoid this issue, prior adaptive methods have limited their use of semi-supervised learning to generating one or a

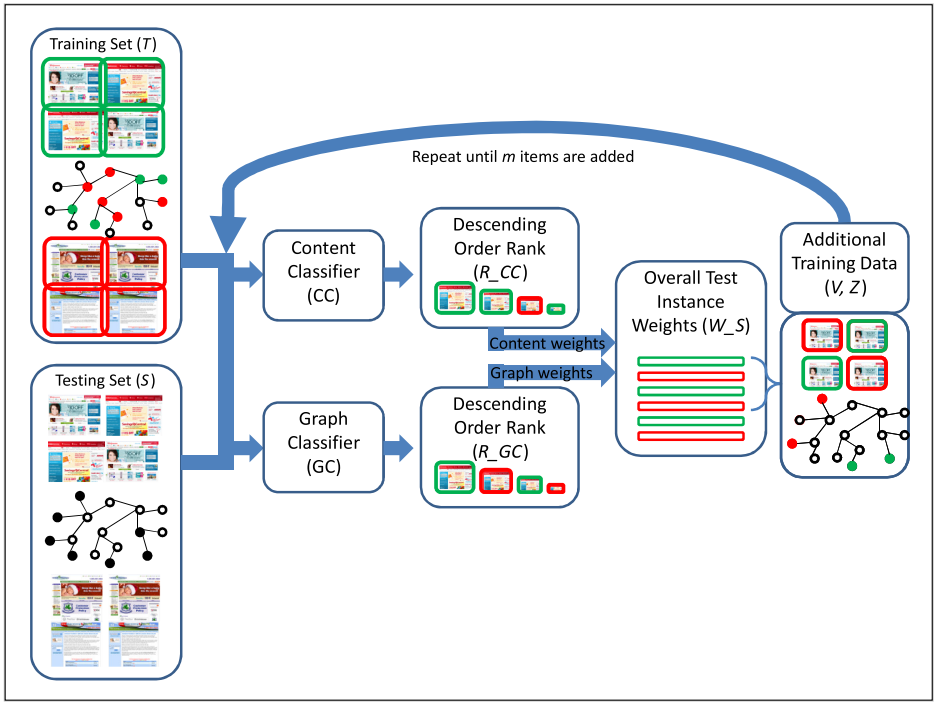

Fig. 2. Recursive trust labeling (RTL) algorithm overview.

few new features [Gan and Suel 2007; Tian et al. 2007]. Consequently, their adaptive learning components only provided marginal improvements over one or two iterations [Castillo et al. 2007]. RTL addresses this issue in two ways. Test instances that have the strongest prediction agreement across the two powerful underlying content and graph-based classifiers are added to the training data. Moreover, during each iteration, the training dataset is reset and all testing instances are reclassified in order to allow error correction. As later demonstrated in the evaluation section, these differences allow RTL to significantly improve performance across several iterations.

Figure 2 shows an overview of the RTL algorithm. A high-level description of RTL's steps is as follows: first, the underlying content and graph-based classifiers are trained and run on the entire test bed. The training set T is then reset to only include the original training instances. The predictions from these classifiers are ranked and these rankings are used to compute an overall weight for each test instance in the starting set S. If the stopping rule has not been satisfied, the d test instances with the highest rank are added to the training data (with class labels congruent with the underlying content and graph classifiers' predictions), and d is incremented. If d is less than the number of instances in the test bed (i.e., m), we repeat all the steps using the expanded training data. Once m items have been added, the algorithm outputs the final test predictions. The details regarding RTL's three main components are discussed below.

4.1. Content Classifier (RTL-CC)

Prior research has noted that site level measures of content-based features occurring across a Web site's pages can be too aggregated, resulting in diminished classification performance [Abbasi and Chen 2009b]. However, since Web sites can contain thousands of pages, incorporating all the content can be computationally inefficient and

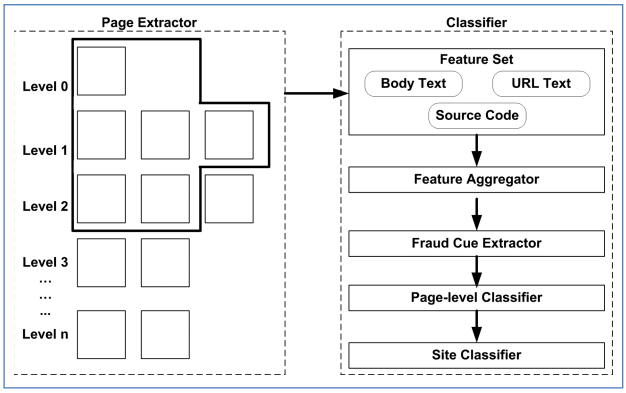

Fig. 3. Illustration of RTL's content classifier.

unnecessary [Ester et al. 2002; Kriegel and Schubert 2004]. Therefore, RTL's content classifier aggregates page-level classification scores from a subset of the Web sites' pages, to derive accurate overall classifications for various Web sites in a computationally efficient manner.

An illustration of the content classifier is presented in Figure 3. For each Web site, the content classifier selects the top x pages. The content feature set utilized is composed of fraud cues derived from the Web pages' body text, source code, and URLs and anchor text. All of these feature categories have been shown to be effective in prior research [Abbasi and Chen 2009b; Drost and Scheffer 2005; Fetterly et al. 2004; Gyongyi et al. 2004; Salvetti and Nicolov 2006]. The body text features used are word n-grams [Drost and Scheffer 2005; Wu and Davison 2006]. Source code n-grams are used for representing page design style [Urvoy et al. 2008; Wu and Davison 2005]. These are extracted from the HTML, JSP, ASP, etc. portions of the page. The URL features include character and token level n-grams extracted from the URLs as well as their anchor text [Araujo and Martinez-Romo 2010; Lin et al. 2007; Salvetti and Nicolov 2006]. For all n-gram feature categories, we use unigrams, bigrams, and trigrams.

Fake medical Web sites often provide inaccurate and/or misleading information [Aphinyanaphongs and Aliferis 2007; Sondhi et al. 2012]. For instance, a fake online pharmacy may intentionally fail to mention the side effects associated with a particular drug in order to bolster sales. Medical terms, including drug names, health conditions, symptoms, adverse reactions, drug-drug interactions etc., often have numerous synonymous or semantically parallel terms and phrases [Liu et al. 2002]. Hence, medical fraud cues may be manifested in many syntactically different (yet semantically equivalent) forms. For instance, clinical drugs have a standardized nomenclature; the psychotropic drug Ciprofloxacin is sold under the brand names Cipro and Proquin. Ciprofloxacin is also known to cause tendonitis or tendon rupture in some patients. Learning fraud cue patterns such as the cooccurrence of "Cipro" + "no side effects" and "Proquin" + "no side effects" would be redundant. Similarly, including patterns based on the cooccurrence of "Cipro" + "tendonitis" and "Cipro" + "tendon inflammation" would also be redundant. This redundancy is problematic since only a finite number of fraud cues can be employed, for computational reasons, to decrease run times, and to avoid over-fitting [Abbasi and Chen 2009b]. Accordingly, we aggregate synonymous concept words using a medical thesaurus, and list of drug names, to remove redundancy and improve fraud cue quality (we used the Unified Medical Language System's MetaThesaurus). For example, all instances of Ciprofloxacin, Cipro, or Proquin in the text are represented using Ciprofloxacin|Cipro|Proquin, while tendonitis and

Let U denote any set of n-grams //e.g., body text unigrams

$u = (u_1, ...u_r)$ denotes a tuple in U

For each u,

$f(u) = H(Y) - H(Y \mid u)$

where $f(u)$ is the weight for feature u, computed using information gain:

$H(Y) = -\sum_{i \in Y} p(Y = i) \log_2 p(Y = i)$ is the entropy across classes Y

$H(Y \mid u) = -\sum_{j \in u} p(u = j) \sum_{i \in Y} p(Y = i \mid u = j) \log_2 p(Y = i \mid u = j)$ is the entropy of $Y \mid u$

Let B denote any higher order set of n-grams //e.g., body text bigrams

For each u, where $f(u) > 0$

Let $C \subseteq B$, where each $c = (c_1, ...c_e)$ denotes a tuple in C with $f(c) > 0$,

and where the tuple u is a part of each c

if $f(u) \geq f(c)$

$f(c) = 0$

Fig. 4. RTL's content classifier's fraud cue extractor.

tendon inflammation would be tendonitis|tendon_inflammation after aggregation. After feature aggregation, all features occurring at least 3 times in the training data are retained.

The formulation for the fraud cue extraction phase is presented in Figure 4. In order to filter out features with lesser discriminatory potential, each feature u is weighted using the information gain heuristic, based on its occurrence distribution across legitimate and fraudulent Web site pages in the training data. While univariate ranking methods such as information gain are computationally effective, they are incapable of removing redundant overlapping n-grams since they consider each feature in isolation [Riloff et al. 2006]. We therefore include a second step in the fraud cue extraction phase where only those higher-order n-grams (i.e., bigrams and trigrams) are retained that have a weight greater than the lower-order n-grams that they are composed of. Once all features have been weighted, a subset of the remaining features (with the highest weights) is incorporated in the final fraud cue set.

The page-level data matrix (composed of the learned features' columns and page instance rows) is input into an SVM classifier that is trained using a linear kernel and the SVM light package [Joachims 2002]. The SVM-generated page-level classification scores on the testing data, which are negative for fake Web sites and positive for legitimate ones, are then input into the site level classifier. Prior work on fake Web site detection has observed that Web pages occurring deeper in the Web site can be more discriminative, since legitimate and fraudulent Web sites' contents often differ more at deeper levels [Abbasi and Chen 2009b]. Accordingly, our site classifier assigns a weight to each page's score that is proportional to its level relative to the home page. A site's classification is based on the weighted sum of its page-level scores, and is presented in Figure 5.

4.2. Graph Classifier (RTL-GC)

Most prior graph classifiers have used single-class propagation (i.e., trust or distrust) along Web site inlinks. The intuition for the latter was based on PageRank principles; fake Web sites can point to legitimate Web sites, but cannot force legitimate Web sites to point to them [Page et al. 1998; Wu and Davison 2005]. However, in the case of fake medical Web sites, this assumption does not always hold true. Given the hefty social

$$CC(A) = \sum_{i=1}^{x} w_i p_i \left(\sum_{i=1}^{x} w_i \right)^{-1}$$

where :

p_i is the page level classifier score for i, one of the x pages extracted from website A

w_i is the weight associated with page i,

which is proportional to v_i, its level relative to A's homepage : $w_i = 1 + \frac{v_i}{l}$

where the parameter l is an integer > 1

Fig. 5. RTL's content-based site classifier (RTL-CC).

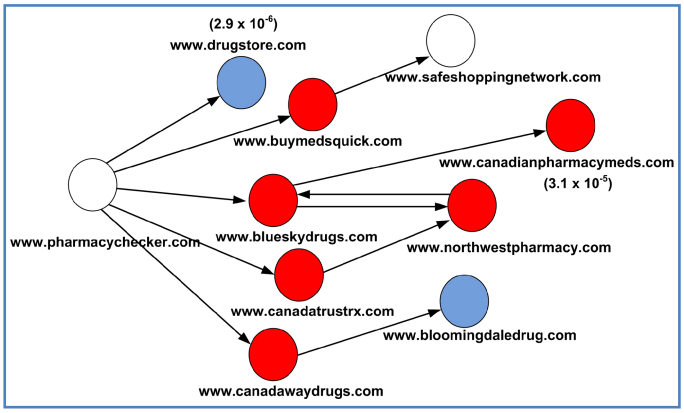

Fig. 6. Illustration of online pharmacy Web site linkage.

costs exacted by fake online pharmacies, there are considerable online communities, discussion forums, and trusted third parties that provide lists of known fake pharmacy Web sites. Moreover, fake pharmacies often include links to other fake Web sites. Figure 6 shows a small illustrative example. In the figure, blue nodes indicate legitimate pharmacies, red denote fraudulent ones, while white nodes indicate other Web sites. The trusted third-party Web site Pharmacy Checker points to numerous legitimate and fraudulent online pharmacies (a few examples are depicted in Figure 6). Moreover, a few of these fake pharmacies point to other legitimate and fake pharmacies. Consequently using TrustRank, a single class inlink-based algorithm, the fake site www.canadianpharmacymeds.com has a considerably higher trust score than the legitimate pharmacy www.drugstore.com (scores are shown next to URLs). While prior dual-class propagation methods quell this concern to a large extent [Wu et al. 2006; Zhang et al. 2006], they typically propagate each class exclusively along inlinks or outlinks (trust based on inlinks and distrust based on outlinks).

We employ a graph classifier that incorporates the aforementioned insights related to fake medical Web sites. Our graph classifier uses dual class propagation simultaneously based on both inlinks and outlinks. Let $S(A)$ denote the initial score of page A, which is "-1" for all known fake pharmacy Web sites, "1" for legitimate ones, and "0" for all other nodes. The overall score $GC(A)$ for a page A is the weighted sum of its propagation score and its initial score, where the propagation score is the weighted normalized sum across A's y inlinks and z outlinks. RTL's graph classifier's mathematical formulation is presented in Figure 7. It is important to note that while RTL's

$$GC(A) = \beta\left(\sum_{i=1}^{y}\left(\alpha\frac{GC(I_i)}{CT(I_i)}\right) + \sum_{j=1}^{z}\left((1-\alpha)\frac{GC(O_j)}{CT(O_j)}\right)\right) + S(A)(1-\beta)$$

where :

$S(A)$ is the initial score of A

$GC(I_i)$ is the GC score of I_i, one of the y inlinks of A

$CT(I_i)$ is the number of outlinks of I_i

$GC(O_j)$ is the GC score of O_j, one of the z outlinks of A

$CT(O_j)$ is the number of inlinks of O_j

α and β are tunable parameters between 0 and 1

Fig. 7. RTL's graph classifier (RTL-GC).

graph classifier embodies many of the same intuitions as the QoC, QoL, and Trust Distrust algorithms [Nie et al. 2007b; Wu et al. 2006; Zhang et al. 2006], it uses a different propagation mechanism. RTL's graph classifier simultaneously propagates trust and distrust based on inlinks and outlinks (see Table IV for formulations of the QoC, QoL, and Trust Distrust algorithms). These differences result in improved fake medical Web site detection performance, as later described in Sections 6.1 and 6.2.

4.3. Recursive Labeling Mechanism

Most prior graph classifiers have used single-class propagation (i.e., trust or distrust) along The mathematical formulation for the recursive labeling mechanism is shown in Figure 8. In each iteration, the content and graph classifiers are run with the training data T and class labels Y. Next, the training dataset T is reset to the original n instances (this is done to allow mislabeled instances to be corrected in future iterations). The predictions for each classifier on the test instances S (i.e., $CC(S)$ and $GC(S)$) are ranked based on their magnitude (i.e., absolute values) in descending order, in R_CC and R_GC. Raw predictions can be positive or negative (where values greater than zero signify that the instance was classified as legitimate). If $CC(i)$ and $GC(i)$ have the same class label, the overall score W_S_i for an instance i is the sum of its two underlying classifiers' ranks, W_CC_i and W_GC_i (where lower scores are better). Otherwise, the score is set to $2m$, where m is the number of testing instances. From W_S, the top d instances are selected (i.e., ones with lowest score) and added to V. The class label Z_{Vi} for a selected instance V_i is assigned based on the polarity and magnitude of the predictions made by $CC(V_i)$ and $GC(V_i)$. Since $CC(V_i)$ and $GC(V_i)$ are not discrete, in situations where the two classifiers make opposing predictions, the prediction with the greater magnitude is used to determine the class label. Finally, V and Z are added to T and Y, respectively. This extended training dataset is used to train the content and graph classifiers in the next iteration. However, since T is reset after its use during each iteration, the $T \cup V$ from iteration 1 would not necessarily be a subset of the $T \cup V$ from iteration 2. The variable d is incremented by a constant p (i.e., $d = d + p$) such that the number of test instances added to the training data increases by p in subsequent iterations. The process is repeated until all testing instances have been labeled (i.e., $d > m$).

5. RESEARCH DESIGN AND EVALUATION TEST BED

Given the gravity of the fake medical Web site problem and specific characteristics exhibited by such Web sites, pertaining to their content and linkage (as mentioned in

Given training examples $T = [t_1, t_2, ... t_n]$, training class labels $Y = [y_1, y_2, ... y_n]$,
and testing instances $S = [s_1, s_2, ... s_m]$
Let $CC(S)$ and $GC(S)$ represent the content and graph classifier predictions using T and Y as input
Initialize variable d to track number of items from S to add to T
where p is a predefined constant and $d = p$
Repeat until $d > m$
  Derive $CC(S)$ and $GC(S)$
  Reset training data to original set of instances $T = [t_1, t_2, ... t_n]$,
    and training class labels $Y = [y_1, y_2, ... y_n]$
  Compute R_CC descending order content prediction rank :

$$R\_CC = [j_1 = \arg\max(e_0 = |CC(S)|), \arg\max(e_1 = e_0(1 : j_1 - 1, j_1 + 1 : m)), ... \arg\max(e_{m-1})]$$

  Compute R_GC descending order graph prediction rank :

$$R\_GC = [j_1 = \arg\max(e_0 = |GC(S)|), \arg\max(e_1 = e_0(1 : j_1 - 1, j_1 + 1 : m)), ... \arg\max(e_{m-1})]$$

  For each test instance i in S, compute the content, graph, and overall weights :

$$W\_CC_i = h, \text{ where } R\_CC_h = i$$

$$W\_GC_i = h, \text{ where } R\_GC_h = i$$

$$W\_S_i = \begin{cases} W\_CC_i + W\_GC_i, \text{ if } CC(i), GC(i) > 0 \text{ or } CC(i), GC(i) \le 0 \\ 2m, \text{ otherwise} \end{cases}$$

  Select the d test instances with the lowest weight (i.e., highest cumulative rank) :

$$V = [j_1 = \arg\min(e_0 = W\_S), \arg\min(e_1 = e_0(1 : j_1 - 1, j_1 + 1 : m)), ... \arg\min(e_{d-1})]$$

  Determine the selected instances' class labels $Z_{V_i} = \begin{cases} 1, \text{if } CC(V_i) + GC(V_i) > 0 \\ -1, \text{if } CC(V_i) + GC(V_i) \le 0 \end{cases}$
  Add selected test instances to training data $T = [t_1, t_2, ... t_n, V]$, $Y = [y_1, y_2, ... y_n, Z]$
  Increment selection quantity variable $d = d + p$
Output T, Y

Fig. 8. Formulation of recursive labeling mechanism in RTL algorithm.

Sections 2 and 4), we conducted experiments to assess the efficacy of RTL's various components. The experiments were designed to answer specific research questions

— *Comparing RTL-CC and RTL-GC against existing content and graph-based methods.* There has been limited prior work pertaining to the detection of fake medical Web sites. *How effective are existing content and graph-based methods for detecting fake medical Web sites?* The proposed RTL-GC is intended to account for linkage tendencies exhibited by medical Web sites. *Can such a graph-based method that employs simultaneous dual-class propagation along in and out links improve performance over existing methods?* Given the complexities associated with medical content, RTL-CC incorporates several novel components (including a medical thesaurus). *Can RTL-CC outperform existing state-of-the-art content-based methods?* Given the challenges associated with identifying fraudulent medical content [Aphinyanaphongs and Aliferis 2007], there is often fairly limited amounts of available training data. *How effective are RTL-CC and RTL-GC, relative to existing content and graph-based methods, when using limited training data?*
— *Comparing RTL against existing meta-learning and adaptive methods.* RTL uses a novel recursive labeling mechanism designed to iteratively add new instances to the training set based on the predictions derived from RTL-CC and RTL-GC.

Table V. Description of Research Test Bed

Domain	Quantity	Sources
Online Pharmacies	166 legit 325 fake	National Association of Boards of Pharmacy (http://www.nabp.net) LegitScript (http://www.legitscript.com)
Medical Institutions	122 legit 22 fake	Hospital Link (http://www.hospitallink.com) Artists Against 4-1-9 (http://wiki.aa419.org/index.php)
Healthcare Information	187 legit 178 fake	Medical Library Association (http://caphis.mlanet.org) Health on the Net (http://www.healthonnet.org) National Council Against Health Fraud (http://www.ncahf.org)

Meta-learning methods such as stacking and adaptive learning strategies have provided state-of-the-art results in prior studies. *How effective is RTL in comparison with existing meta-learning methods or adaptive techniques (which typically use semi-supervised learning to iteratively generate new features)?* Moreover, considering that RTL's recursive labeling mechanism is designed to effectively leverage unlabeled instances: *How robust is RTL when using limited quantities of balanced and imbalanced training data?*

—*Assessing the impact of medical Web site categories on the performance of RTL and state-of-the-art methods.* The various categories of fake medical Web sites each have different objectives and characteristics. *How effective are RTL and state-of-the-art comparison methods for detection of fake pharmacy, medical institution, and health information Web sites?*

In the remainder of Section 5, we describe the experimental test bed, encompassing hundreds of known legitimate and fake medical Web sites. Section 6 includes a detailed evaluation of RTL-CC, RTL-GC, and RTL as a whole, in comparison with numerous content, graph, metalearning, and adaptive methods.

5.1. Test Bed

The test bed comprised 1,000 legitimate and fake medical Web sites (described in Table V). We collected 166 legitimate and 325 fraudulent online pharmacy Web sites. The URLs for these Web sites were obtained from reputable sources including the National Association of Boards of Pharmacy and LegitScript, both of which follow rigorous, well-documented standards for evaluating online pharmacies. The number of fraudulent instances was nearly double since fake pharmacies exceed legitimate ones [Easton 2007; Greenberg 2008]. Our own analysis of Google search query results for "online pharmacy" in October 2009 and December 2012 revealed that 61 and 58 of the top 100 search results were fraudulent, respectively. We also collected 122 legitimate and 22 fake medical institution Web sites, primarily composed of legitimate and fake hospitals. The legitimate URLs were taken from Hospital Link, an online portal that contains a list of trusted hospitals. The fake hospitals' URLs were identified through the Artists Against 4-1-9 (an online fraud prevention community), as well as various news articles appearing in online outlets such as CNN and BBC. Due to the lack of online databases specifically for fake hospital Web sites, and the fact that they are often taken down very quickly, the number collected was relatively smaller. Furthermore, we collected 187 legitimate and 178 fraudulent health information Web sites. The legitimate URLs were those taken from the Medical Library Association's consumer and patient health information section (CAPHIS), as well as ones certified by the Health on the Net Foundation [Gaudinat et al. 2007; Wang and Richard 2007]. The fraudulent health information Web sites were taken from the National Council Against Health

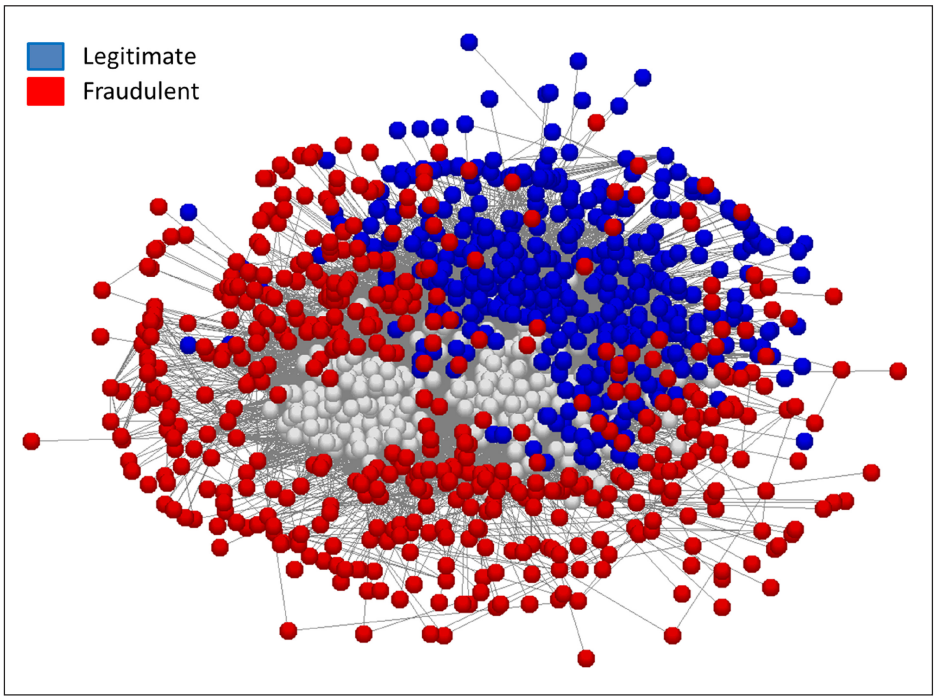

Fig. 9. Subset of the site-level test bed graph.

Fraud's (NCAHF) Web site and discussion list. The NCAHF discussion list includes several physicians and medical researchers that post URLs for questionable health information Web sites (e.g., ones claiming that autism is a curable virus). Members of the list discuss the merits of the claims made in each of the Web sites. We included only Web sites where there was an overwhelming consensus amongst the discussants. The NCAHF has also been used as a resource in previous research [Aphinyanaphongs and Aliferis 2007].

All pages within each Web site were collected leading to an initial collection of approximately 2.7 million pages. Surprisingly, the fake medical Web sites were much larger than the legitimate ones, with nearly twice as many pages. The entire hyperlink graph for these 1,000 seed URLs was collected using link expansion. This involved iteratively collecting all inlinks and outlinks of queue pages, adding the newly collected pages to the queue, and repeating the process. The graph encompassed nearly 18 million pages from approximately 930,000 Web sites/domains, with close to 100 million links. Figure 9 shows a small subset of the site level graph. Depicted are the 1,000 known nodes as well as additional nodes with a degree greater than 1,000. Legitimate medical Web sites are blue while known fakes are red and other nodes are colored white. The graph layout was determined using a spring-embedded algorithm. Visually, it appears that the application of graph-based methods may be feasible as both legitimate and fraudulent sites appear to be situated in close proximity to others belonging to the same class.

6. EVALUATION

A bootstrapping approach was employed for all experiments, where the 1,000 Web sites were randomly split into 150 training and 850 test cases, for 30 independent

runs. In each of the 30 bootstrap runs, the training data was composed of 75 legitimate and 75 fraudulent instances, while the testing data contained 400 legitimate and 450 fraudulent instances. All analysis was performed at the site level, as done in many prior studies [Gyongyi et al. 2006; Wu and Chellapilla 2007]. Our preliminary analysis revealed that the graph-based methods performed better on site-level graphs as compared to page-level ones. Hence, for each bootstrap run, the algorithms were only evaluated on the 850 test sites (though the graph-based methods were run on the entire 930,000 node graph). The evaluation metrics employed included overall accuracy and class-level precision, recall, and f-measure [Wu and Davison 2005]. Three different sets of experiments were run. In the first set, RTL's content and graph classifiers (i.e., RTL-CC and RTL-GC) were compared against various content and graph-based methods for fake Web site detection. In the second, RTL was evaluated in comparison with several stack classifiers. Experiment three compared RTL against existing adaptive methods.

6.1. Results for RTL-CC and RTL-GC in Comparison with Content and Graph-Based Methods

We compared RTL-CC and RTL-GC against 9 content and 10 graph-based methods. Each of the 19 comparison algorithm's parameters were tuned extensively; the settings yielding the best results were utilized. The graph-based algorithms utilized were the ones described in Tables III and IV: PageRank, TrustRank, AntiTrustRank, BadRank, Mass Estimation, Parent Penalty, QoC, QoL, TrustDistrust, and Cautious Surfer. In the case of the graph-based methods, an additional classification threshold parameter s was also used. Sites were classified as legitimate/fake depending on their position with respect to the threshold, which was tuned for each graph-based algorithm.

BadRank was run using $d = 0.5$ and $s = 0.001$, while $d = 0.1$ and $s = 0.000001$ were used for AntiTrustRank. For ParentPenalty, $t = 3$ and $p = 4$. TrustRank was run using $d = 0.5$ and $s = 0.000001$ while these same parameters were set to 0.1 and 0.999999 for Mass Estimation, respectively. Cautious Surfer was run using $s = 0.03$. QoC and QoL were run with $\alpha = 0.5$, $\beta = 0.5$, $d = 0.85$, and $s = 0$ while these same parameters were set to 0.5, 0.7, 0.1, and 0.0000005 for TrustDistrust, respectively. For PageRank, d was set to 0.05, and 0.0000001 was used for s.

As noted in Section 3.1, prior content-based detection studies commonly utilized machine learning classifiers in conjunction with n-gram features derived from body text, URL and anchor text, and source code. Accordingly, the comparison content-based classifiers employed were ones utilized in prior studies: linear kernel SVM, polynomial kernel SVM, RBF kernel SVM, Bayesian Network, Naïve Bayes, Neural Network, C4.5 Decision Tree, and Logistic Regression [Abbasi and Chen 2009a, 2009b; Drost and Scheffer 2005; Kolari et al. 2006; Ntoulas et al. 2006]. Each classifier was run using body text, URL and anchor text, and HTML n-grams. Since the number of content features employed in prior studies has varied considerably, each classifier was run using the top 2,500 to 15,000 features (ranked based on their information gain weights) in 2,500 feature increments. For each of the 8 classifiers, we used the feature quantity which yielded the best performance results on the testing data. The Neural network was run using 2,500 features. The Bayesian Network was run using 5,000 features. C4.5, Logistic Regression, and Naïve Bayes were all run using 7,500 features. SVM-RBF was run using 7,500 features with gamma = 0.001. SVM-Poly was run using 10,000 features with $d = 2.0$. SVM-linear was run using 12,500 features. All of the aforementioned content-based classifiers were run using either Weka or SVMLight [Joachims 2002; Witten and Frank 2005]. We also included AZProtect as a benchmark content-based detection method since it had performed well in previous

Table VI. Results for RTL-GC and RTL-CC in Comparison with Content and Graph Methods

	Algorithm	Overall Accuracy	Legit			Fake		
			F-meas.	Prec.	Recall	F-meas.	Prec.	Recall
Graph	RTL-GC	89.24	87.84	93.78	82.60	90.35	86.02	95.13
	QoC	86.39	84.50	91.02	78.85	87.86	83.19	92.38
	Mass Estim.	83.22	80.77	87.64	74.91	85.11	80.25	90.61
	QoL	82.73	80.70	85.14	76.70	84.38	80.96	88.10
	TrustDistrust	82.37	79.01	89.84	70.51	84.80	78.00	92.91
	TrustRank	78.92	76.55	80.34	73.10	80.86	77.86	84.10
	AntiTrustRank	70.58	66.52	71.61	62.10	73.76	69.87	78.11
	BadRank	66.62	69.68	60.85	81.49	62.88	76.45	53.40
	Cautious Surf	63.55	58.10	63.28	53.70	67.74	63.73	72.30
	PageRank	58.70	68.60	53.41	95.87	39.67	87.48	25.65
	ParentPenalty	51.98	49.79	49.02	50.58	54.00	54.79	53.23
Content	RTL-CC	91.02	90.05	94.14	86.30	91.82	88.66	95.22
	AZProtect	88.56	87.26	91.73	83.20	89.63	86.21	93.33
	SVM-Linear	86.62	84.99	90.00	80.50	87.93	84.16	91.89
	Logistic Reg.	83.84	82.21	85.29	79.34	85.20	82.71	87.84
	SVM-RBF	82.40	80.15	85.38	75.53	84.18	80.27	88.50
	SVM-Poly	80.80	78.76	82.13	75.66	82.48	79.78	85.37
	Bayes Net	79.77	78.14	79.50	76.82	81.17	79.99	82.39
	Neural Net	79.44	77.08	81.08	73.46	81.36	78.23	84.76
	C4.5	78.75	76.60	79.48	73.92	80.53	78.18	83.04
	Naïve Bayes	78.56	74.08	85.96	65.08	81.73	74.47	90.55

studies [Abbasi and Chen 2009a, 2010]. AZProtect was run using a rich feature set composed of nearly 6,000 attributes derived from the Web sites' body text, source code, URL tokens, images, and linkage-based information [Abbasi and Chen 2009a]. These features were learned from the training Web sites (in each bootstrap fold), using the information gain heuristic. RTL-GC and RTL-CC's parameters were tuned using 10 fold cross-validation on the training data. RTL-CC was run using 100 pages per site (x = 100). RTL-GC was run using α = 0.5, β = 0.5, and s = 0.0001. In order to allow a fair comparison against the content and graph methods, RTL-CC and RTL-GC were run as stand-alone methods, without the use of the recursive labeling mechanism.

Table VI shows the experimental results, averaged across the 30 bootstrap runs. RTL-GC and RTL-CC outperformed all their respective comparison graph and content classifiers across all seven evaluation metrics. With respect to the comparison techniques, graph-based methods such as QoC and Mass Estimation and content-based methods such as AZProtect and SVM-Linear had the best performance. As suspected, PageRank performed poorly since it is highly susceptible to exploitation via link farms. It is worth noting that three of the best comparison graph-based methods (i.e., QoC, QoL, and TrustDistrust) were the only three that used dual class propagation (i.e., they propagated trust and distrust). Their improved performance is consistent with prior research, which has alluded to the superiority of propagating trust and distrust over simply using single class propagation [Wu et al. 2006]. However, as noted in Section 4.2 (and illustrated in Figure 6), in the case of medical Web sites, simultaneously propagating trust and distrust along both in and out-links can yield better results. This allowed RTL-GC to outperform existing dual class propagation methods. The enhanced performance of AZProtect over other comparison content-based methods is in line with prior concocted Web site detection studies [Abbasi and Chen 2009b; Abbasi et al. 2010]. With respect to the comparison methods, the content-based

techniques not only had the best overall results, they were also more consistent in their performance (i.e., they had less variation).

Pairwise t-tests were conducted on the accuracy and class-level f-measure values for the 30 bootstrap runs (n = 30). In the t-tests, RTL-CC was evaluated against the 9 comparison content classifiers, while RTL-GC was compared against the 10 graph classifiers. RTL-GC and RTL-CC significantly outperformed their comparison methods (all p-values <0.001, n = 30), with overall accuracy values that were at least 2% higher. The results suggest that RTL-GC and RTL-CC better leveraged and exploited key characteristics of fake medical Web sites, enabling enhanced performance over comparison methods.

6.2. Robustness Analysis of RTL-CC and RTL-GC in Comparison with Content and Graph-Based Methods

Since often there is limited knowledge of known good or bad sites [Gyongyi et al. 2004; Wu and Chellapilla 2007], algorithms that are less dependent on larger training datasets are highly desirable. Accordingly, we assessed the robustness of RTL-CC, RTL-GC, and the comparison algorithms. Here, robustness refers to the effectiveness of the algorithms when run using less training data. Each algorithm was run using 10%−100% of the original training data (in 10% increments), for each of the 30 bootstrap runs. In other words, each algorithm was run using between 15 and 150 training instances, in 15 site increments. For each bootstrap run, all testing instances were used, as done in the previous experiment (Section 5.1). Since PageRank is an unsupervised method that does not rely on training data, changes to the training data did not impact its results. It was therefore excluded from the analysis.

Figure 10 shows the experimental results across the 30 bootstrap runs. The figure depicts charts for RTL-GC and RTL-CC in comparison with the content and graph-based methods. The charts include overall accuracy (top row) and class-level f-measures, for different training set sizes. For each chart, the y-axis indicates the algorithms' accuracy or f-measure while the x-axis displays the percentage of the original training data used. As shown in the figure, RTL-GC outperformed all comparison graph methods, while RTL-CC outperformed the content-based methods for all test bed sizes. RTL-GC and RTL-CC were both more effective than comparison methods when using less training data. In the case of RTL-GC, this was attributable to the use of simultaneous dual class propagation along both in and out-links, which allowed the more accurate propagation of trust and distrust with sparse training data. In contrast, the best content and graph-based methods needed at least 40%–50% of the training data (e.g., AZProtect, SVM-linear, Logit, QoC, Mass Estimation) while some required even more (e.g., BadRank, AntiTrustRank), to attain results comparable to those yielded using the entire training set. Case in point, QoC and AZProtect's accuracies were 6% and 10% lower when using only 20% of the training data, respectively.

Table VII shows the area under the curve (AUC) values for RTL-GC and RTL-CC in comparison with the content and graph-based algorithms. The AUC values are based on the overall accuracy and class-level f-measure plots depicted in Figure 10, as well as the legitimate and fake precision/recall plots. Based on the table, it is further evident that RTL-GC and RTL-CC outperformed comparison classifiers in terms of AUC values associated with overall accuracy and class-level f-measures, precision, and recall. These results support the notion that RTL's underlying content and graph classifiers can dramatically improve performance over existing methods: the RTL-GC and RTL-CC AUC values for accuracy and f-measures were at least 30 to 55 points higher than those associated with the best graph and content methods.

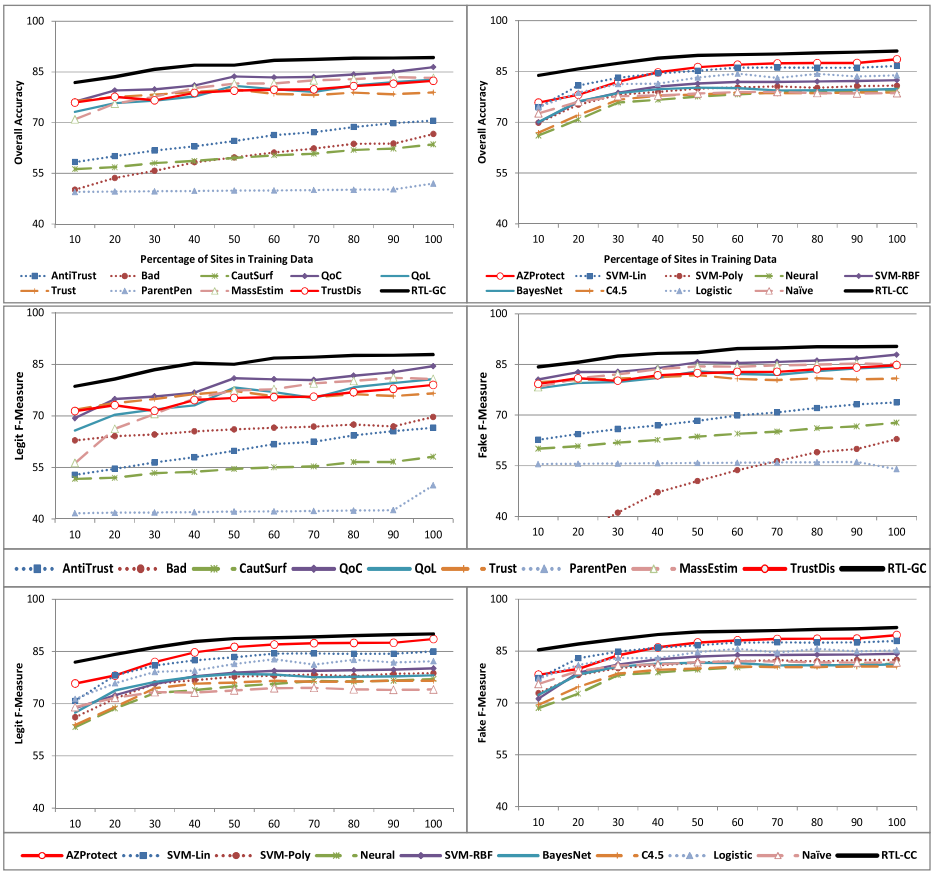

Fig. 10. Robustness results for RTL-CC and RTL-GC in comparison with content and graph-based methods.

6.3. Results for RTL in Comparison with Stack Classifiers

In the second set of experiments, RTL was compared against stacked classifiers. Stacking is a popular metalearning strategy that has worked well in prior fake Web site detection studies [Abbasi and Chen 2009b; Becchetti et al. 2008]. RTL was run using both underlying classifiers (i.e., RTL-CC and RTl-GC) in conjunction with the recursive labeling mechanism. The recursive labeling mechanism was run using p = 50 since this setting provided a good balance between run time and performance when using 10-fold cross validation on the training data. Eight of the content-based classifiers from the prior experiment were used as top-level classifiers in the stack. These stack classifiers made predictions using the underlying content and graph-based classifiers' Web site predictions/scores as their input feature values. As a preprocessing step, we sorted all 19 content and graph classifiers based on their performance in the prior experiment and added them one at a time as features for the 8 stacks. Hence, AZProtect was added first, followed by SVM-linear, QoC, Logistic regression, Mass Estimation, QoL, etc. Given the poor performance of certain classifiers, this approach was utilized in order to allow the stack classifiers to use the less noisy underlying classifiers as features, thereby improving their performance. The results from this preprocessing step are presented in Figure 11. Generally, using the top 2 or

Table VII. AUC Values for RTL-GC and RTL-CC and Comparison Methods

	Algorithm	Overall Accuracy	Legit			Fake		
			F-meas.	Prec.	Recall	F-meas.	Prec.	Recall
Graph	RTL-GC	784.10	767.03	831.19	712.40	797.25	752.52	847.83
	QoC	741.37	710.78	805.45	637.40	763.26	704.26	833.42
	Mass Estim.	722.53	676.29	818.04	584.97	751.84	679.64	844.81
	TrustDistrust	713.44	675.37	778.96	596.17	740.46	676.60	817.68
	QoL	710.60	676.68	755.83	614.29	735.29	683.77	796.21
	TrustRank	706.31	679.70	730.86	635.75	727.09	689.74	769.03
	AntiTrustRank	585.80	542.57	584.03	506.61	619.70	587.06	656.19
	Cautious Surf	538.28	491.88	525.08	462.79	575.13	547.86	605.38
	BadRank	536.70	594.28	495.19	747.35	445.82	642.29	349.47
	ParentPenalty	450.11	382.92	357.30	412.71	501.56	531.11	475.20
Content	RTL-CC	800.50	790.44	819.56	763.36	808.87	785.67	833.51
	AZProtect	762.55	747.67	781.20	716.99	774.76	748.46	803.04
	SVM-Linear	758.48	740.29	788.46	697.71	772.95	737.09	812.42
	Logistic Reg.	738.94	721.17	754.13	691.10	753.48	727.49	781.47
	SVM-RBF	720.87	697.53	747.28	654.48	739.24	702.95	779.89
	SVM-Poly	709.18	687.06	722.50	655.10	727.09	699.35	757.25
	Bayes Net	708.00	690.71	708.89	673.54	722.62	707.36	738.63
	Naïve Bayes	700.61	660.65	760.90	584.09	729.07	666.94	804.19
	C4.5	692.17	671.39	695.95	648.53	709.47	689.24	730.95
	Neural Net	688.49	665.99	696.50	638.25	706.99	682.77	733.16

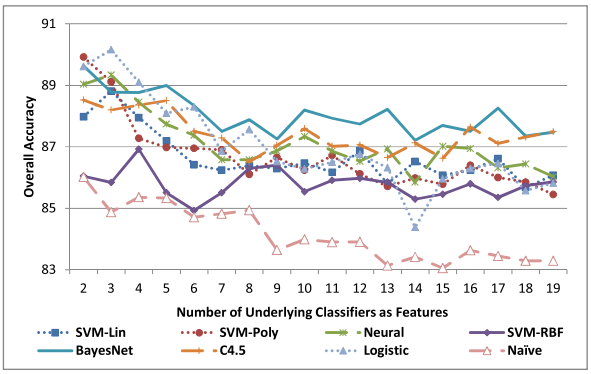

Fig. 11. Preprocessing step results for comparison stack classifiers.

3 underlying classifiers as features provided the best performance for the stacks. In the case of SVM-RBF, four base classifiers were employed (i.e., AZProtect, SVM-linear, QoC, and Logistic regression). The best setting for each of the 8 stack classifiers was compared against RTL in the ensuing experiments.

Table VIII shows the experimental results across the 30 bootstrap runs. RTL outperformed all 8 comparison stack classifiers in terms of overall accuracy and class-level f-measures, precision, and recall. RTL's performance gain over the comparison stack classifiers was between 4% and 8% in terms of overall accuracy. While RTL outperformed the comparison stack classifiers on legitimate and fake medical Web sites, the performance gain was particularly large on the legitimate Web sites (as evidenced by the class-level recall values). In other words, RTL had considerably lower false positive

Table VIII. Results for RTL in Comparison with Content and Graph-Based Methods

Algorithm		Overall Accuracy	Legit			Fake		
			F-meas.	Prec.	Recall	F-meas.	Prec.	Recall
RTL		94.33	93.83	96.22	91.55	94.76	92.80	96.80
Stack	Logistic Reg.	90.16	88.94	94.44	84.05	91.14	87.09	95.60
	SVM-Poly	89.92	89.00	91.47	86.66	90.70	88.68	92.82
	Bayes Net	89.63	88.31	94.06	83.22	90.69	86.47	95.33
	Neural Net	89.34	88.30	91.34	85.45	90.21	87.77	92.80
	SVM-Linear	88.81	87.65	91.22	84.34	89.77	86.96	92.78
	C4.5	88.52	87.25	91.38	83.47	89.56	86.36	93.00
	SVM-RBF	86.92	85.84	87.53	84.22	87.86	86.43	89.33
	Naïve Bayes	86.02	84.57	87.94	81.45	87.21	84.53	90.07

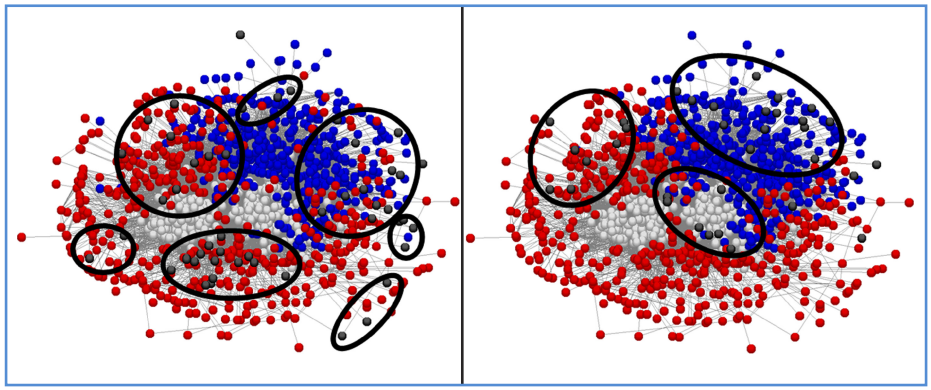

Fig. 12. Classification graphs for content (left) and graph-based (right) method.

rates (with 5% to 10% higher recall on legitimate Web sites), whereas the comparison methods were prone to misclassifying legitimate Web sites as fake. With respect to the comparison stack classifiers, Logistic Regression and SVM-Polynomial had the best performance, followed by Bayesian Network and Neural Net. Considering that the stack classifiers were retrospectively tuned for optimal performance (see Figure 11), the results especially underscore RTL's enhanced effectiveness over comparison stack classifiers.

Pairwise t-tests were conducted on the accuracy values for the 30 bootstrap runs (n = 30). The t-tests compared RTL against the 8 stack classifiers. For all 8 t-tests, RTL significantly outperformed the stack classifiers (all p-values < 0.001). The results suggest that the effectiveness of RTL was partially attributable to the complementary information provided by its underlying RTL-CC and RTL-GC methods. Figure 12 shows the classification graphs for RTL-CC (left side) and RTL-GC (right side) from a bootstrap run where RTL attained very good results. The blue (legitimate) and red (fake) nodes indicate training sites or correctly classified testing sites. The misclassified sites are colored black and contained within the black ovals. Both RTL-CC and RTL-GC misclassified approximately 4% of the test Web sites. However, the two sets of misclassifications were mutually exclusive. RTL-GC's misclassifications (depicted in the right side of Figure 12) came from three regions where legitimate Web sites were positioned in areas predominantly composed of fake sites, and vice versa. In contrast, RTL-CC's misclassifications (which were based on content-based similarities) came from various regions of the graph; many of these sites were correctly classified

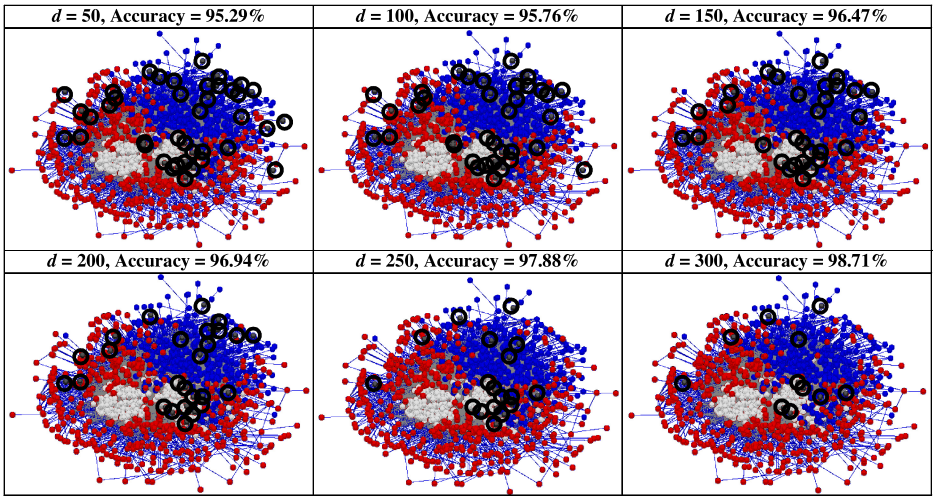

Fig. 13. Illustration of RTL's recursive labeling mechanism.

by the graph classifier. By using predictions from both classifiers in unison, RTL was able to correctly classify over 98% of the sites for this particular bootstrap run.

Another critical component of RTL was its use of recursive labeling. Figure 13 illustrates how RTL was able to gradually improve its performance by recursively relabeling test instances as additional information became available. The figure shows RTL's performance through six iterations (i.e., $d = 50$ to $d = 300$) for a particular bootstrap run. As with the previous figure, blue and red nodes indicate training sites or correctly classified testing sites, while the misclassified sites are colored black and contained within black circles. Initially, there were approximately forty misclassified instances. However, the number decreased during subsequent iterations, until only eleven instances were eventually misclassified. Collectively, Figures 12 and 13 exemplify how RTL's use of complementary information from powerful underlying content and graph classifiers, and its recursive labeling mechanism facilitate the enhanced detection of fake Web sites.

6.4. Robustness Analysis of RTL in Comparison with Stack Classifiers

Robustness analysis was conducted for RTL in comparison with the stack classifiers. The experimental setup employed was the same as in the prior robustness experiment (Section 6.2): each algorithm was run using 10%−100% of the original training data (in 10% increments), for each of the 30 bootstrap runs. Table IX shows the AUC values for RTL and the 8 comparison stack classifiers. RTL had better overall accuracy and class-level f-measures than the 8 stacks. It outperformed the comparison stacks by 67–134 points in terms of AUC values for overall accuracy. These gains were evident on legitimate and fake Web sites; RTL's class-level f-measures were at least 59−78 points higher than the best stack classifier. With respect to the comparison methods, Logistic regression and SVM-Poly had the best performance.

Figure 14 shows the overall accuracy and class-level f-measure graphs for the 8 stack classifiers, compared against RTL. The y-axes indicate accuracy while the x-axes display the percentage of the original training data used. Based on the overall accuracy performance gaps, it is apparent that RTL outperformed all of the stack classifiers, and generally by a wide margin. This improvement was largely due to better performance on legitimate Web sites, where RTL had f-measures that were consistently at

Table IX. AUC Values for RTL in Comparison with Stack Classifiers

Algorithm		Overall Accuracy	Legit			Fake		
			F-meas.	Prec.	Recall	F-meas.	Prec.	Recall
RTL		837.22	831.57	852.67	811.52	842.00	824.72	860.06
Stack	Logistic Reg.	770.46	753.44	803.93	708.98	783.93	746.71	825.11
	SVM-Poly	770.18	758.78	777.51	740.96	779.87	764.26	796.15
	Bayes Net	762.13	743.21	797.21	696.12	776.97	737.61	820.80
	SVM-Linear	736.59	720.34	748.72	694.11	750.13	727.43	774.36
	Neural Net	722.03	705.12	729.24	682.63	736.22	716.58	757.06
	SVM-RBF	717.01	679.73	667.88	692.88	743.28	753.98	733.41
	C4.5	716.16	703.38	712.53	694.69	727.33	719.77	735.24
	Naïve Bayes	703.55	652.78	662.44	643.73	736.96	730.77	743.43

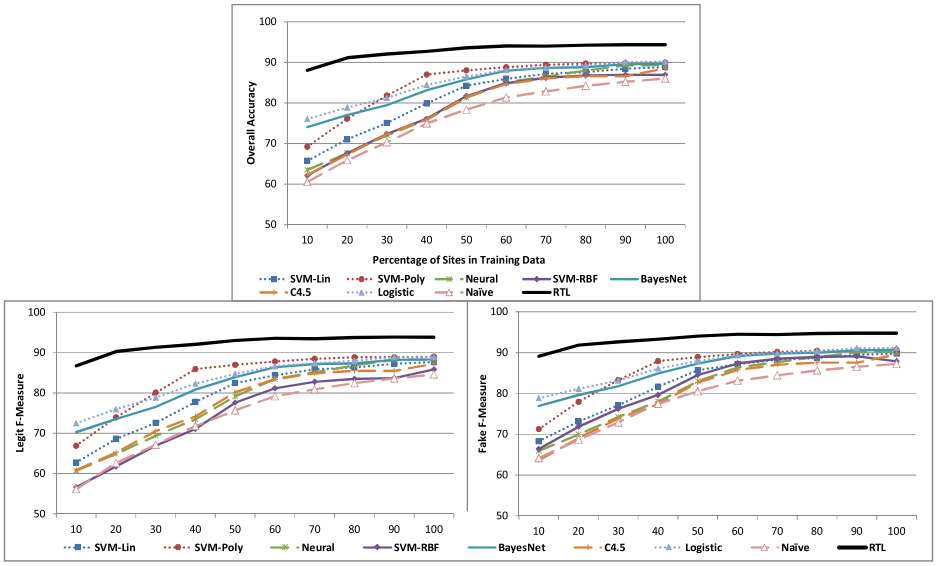

Fig. 14. Robustness results for RTL (thick solid) and stack classifiers.

least 5%−14% higher than the best comparison stacks. As previously alluded to, RTL's improved recall on legitimate Web sites resulted in lower false positive rates. RTL's enhanced performance was also attributable to its ability to attain considerably better performance when using very little training data (i.e., 10%–20%). It attained 91.1% accuracy when using only 30 training sites. This suggests that the adaptive learning mechanism used by RTL is highly effective at accurately classifying legitimate and fake medical Web sites even when the number of known good and bad instances is very small. In contrast, the stack classifiers performed very poorly when the amount of available training data was limited. For instance, SVM-Linear, SVM-RBF, Neural Network, and Naïve Bayes all had accuracies close to 60% when using 10% of the training data. Even the best stack classifiers (Logistic regression and SVM-Polynomial) had overall accuracies of 69% and 76% respectively, when using 10% of the training data. Conversely, RTL had approximately 88% accuracy when using the same quantity of training instances. Hence, RTL's performance gain was larger for smaller training set sizes; it outperformed all comparison methods by at least 10% in terms of overall accuracy when using 10%−20% of the training data.

6.5. Results for RTL in Comparison with Adaptive Methods

We evaluated the effectiveness of RTL in comparison with three existing adaptive methods: Combinatorial Feature-Fusion [Tian et al. 2007], Secondary Classifier Approach [Gan and Suel 2007], and Stacked Graphical Learning [Castillo et al. 2007]. For Combinatorial Feature-Fusion (CFF), we first extracted 207 features for each Web site in the dataset as suggested by Tian et al. [2007]. Initially, each bag-of-word in the training data was weighted using the authors' proposed rank performance weighting method. The 200 bag-of-words with the highest weight were included in the feature set. Also included were seven "engineered" features, five of which were as follows: the number of inbound/outbound links to/from this site, the fraction of total links that were inbound/outbound, and a binary feature that was "1" if any of the bag-of-words' tf-idf values were above 0.2. The two remaining "engineered" features were the percentage of the 500 hotwords (most frequently occurring words in the test bed) covered by the site and the percentage of total unique words in a site that were hotwords. These 207 features were used to train a rough ADTree classifier [Anderson and Moore 1998] that was applied to all the nontraining site nodes in the graph. Once each site in the graph had a label (where the training sites had their actual label while the remainder had the one predicted by the "rough" classifier), four additional features were generated for each graph node. These were the number of inbound/outbound fake links and the percentage of inbound/outbound fake links. These features were added to the original 207, and the resulting 211 features were used to retrain and rerun the rough classifier on all the nontraining sites. After this second round of rough classification, the four graph-based features were recomputed for each node. This final set of 211 features were fused (in pairs of two) and a subset of these fused features were employed in the final ADTree model, following the approach taken by Tian et al. [2007].

Gan and Suel [2007] evaluated several relabeling methods for improving the performance of an underlying classifier and found the Secondary Classifier Approach (SCA) to be the most effective. Following the SCA approach, we initially trained a basic classifier [Gan and Suel 2007]. A C4.5 decision tree model [Quinlan 1986] was built using 25 features which included 8 content, 14 link, and 3 domain registration-based attributes. The basic classifier was applied to all the nontraining nodes in the graph. A secondary C4.5 classifier was then trained using seven features for each site, which included: the basic classifier's prediction label and confidence score, the percentage of incoming/outgoing links from/to fakes, and the fraction of weighted fakes in the incoming/outgoing neighbors (where the neighbor nodes' weights were based on their basic classifications' confidence scores). This secondary classifier was used to assign a final label to each node.

For Stacked Graphical Learning (SGL), we initially trained a base C4.5 classifier [Castillo et al. 2007]. The classifier used numerous link and content-based features adapted from Becchetti et al. [2006] and Ntoulas et al. [2006]. The link-based features included degree-based measures and ones derived using PageRank and TrustRank, amongst others. The content-based features included average word length, number of words in the page title, amount of anchor text, fraction of visible content, n-gram occurrence likelihoods, etc. The base classifier was run on all nontraining nodes in the graph. For each node in the data, an additional feature was computed: the fraction of all in/out-link nodes that were classified as fake. This new feature was added to the original feature set, which was used to retrain and rerun the base classifier. The process was repeated for several iterations and the results from the iteration yielding the best performance were reported. In the ensuing experiments, the best results were consistently attained when running SGL for two iterations (same as Castillo et al. [2007]).

Table X shows the average experimental results across the 30 bootstrap runs. RTL outperformed the three comparison adaptive methods across all seven evaluation

Table X. Results for RTL in Comparison with Adaptive Methods

Algorithm	Overall Accuracy	Legit			Fake		
		F-meas.	Prec.	Recall	F-meas.	Prec.	Recall
RTL	94.33	93.83	96.22	91.55	94.76	92.80	96.80
SGL	90.59	89.51	94.07	85.38	91.46	87.99	95.22
CFF	89.56	88.70	90.42	87.05	90.30	88.86	91.80
SCA	87.02	85.40	90.71	80.68	88.31	84.37	92.65

metrics (all pair–wise t-test p-values were significant at alpha of 0.05). The performance gains in terms of overall accuracy and class-level f-measures were between 3.3% and 8.4%. Interestingly, the best stack classifier (Logit Regression from Section 6.3) outperformed two of the three comparison methods. This result suggests that the information from powerful graph-based classifiers may serve as better input features for secondary classifiers than the relatively simpler link-based attributes adopted by certain adaptive methods. The Logit stack used QoC node scores as an input feature for enhanced performance. It is therefore no coincidence that SGL, the adaptive method that attained the best results, used various measures derived from the PageRank and TrustRank algorithms as input features. Similarly, RTL-GC played an integral role in RTL's enhanced performance, as illustrated in Figure 12. Another important factor, which was alluded to in Section 4.3, was that all three comparison adaptive methods used semi-supervised learning to generate new features, which were added to the training data. This caused the relabeling process to plateau after 1–2 iterations, as noted in previous studies and also observed in this experiment. In contrast, RTL's recursive labeling mechanism actually adds new instances to the training models, thereby allowing performance to improve across several iterations (as illustrated in Figure 13).

Robustness analysis was conducted for RTL in comparison with the adaptive methods. Two different experimental settings were used. The first was similar to the setup employed in the prior robustness experiments: each algorithm was run using 10%–100% of the original training data (in 10% increments), for each of the 30 bootstrap runs. In the second setting, we analyzed the impact of class imbalance on the performance of the online pharmacy and medical institution segments of the test bed. This was done by varying the quantity of minority class instances in the training data from 5% to 100% in increments of 5%, while holding the majority class instances constant. In the case of online pharmacies, since fraudulent pharmacies outnumber legitimate ones and therefore constitute the majority class, we kept the quantity of fraudulent training instances constant while fluctuating the number of legitimate instances. In contrast, the number of legitimate medical institution sites was held constant while the number of fraudulent ones was varied. This setting was included to assess the impact of using imbalanced training data on the performance of adaptive methods; since an equivalent quantity of known legitimate/fake medical instances may not always be readily available [Aphinyanaphongs and Aliferis 2007]. For this imbalanced setting, receiver operating characteristic (ROC) curves were generated in order to show the tradeoffs between true positive and false positive rates.

Figure 15 displays the robustness results using the standard setup (tops) as well as the imbalanced setting (bottom). For the balanced setting, RTL outperformed all three comparison methods in terms of both legit and fake recall. The performance margins appeared to increase as the quantity of training data decreased. While using imbalanced training data, RTL's ROC curves dominated the three comparison methods on both datasets, suggesting that RTL provided better ratios of true positives to false positives than the three comparison adaptive methods across different levels of

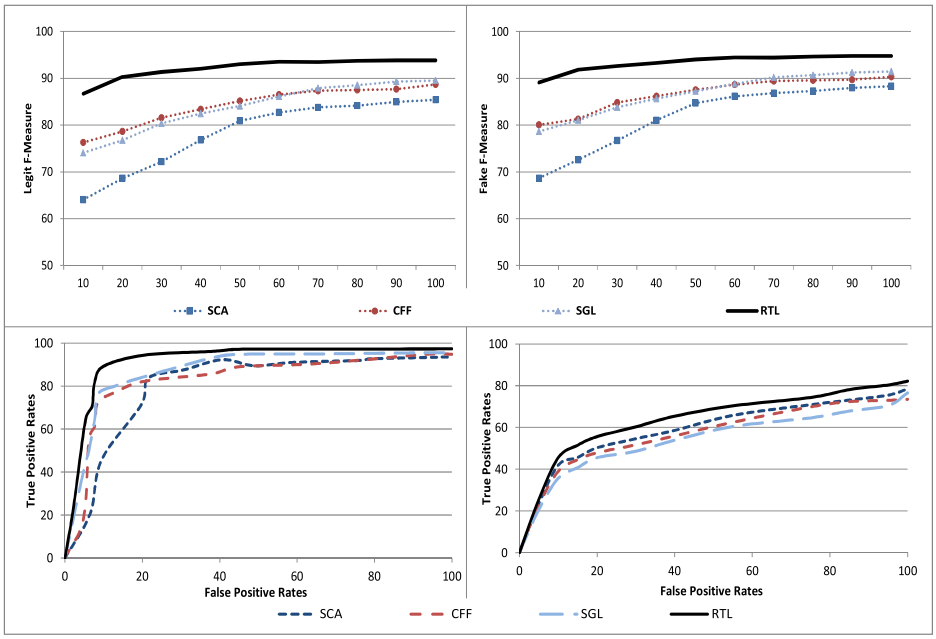

Fig. 15. Robustness results for RTL and adaptive methods: Legit and fake F-measures using balanced training data (top); ROC curves using imbalanced training data for online pharmacies (bottom left) and medical institutions (bottom right).

imbalanced pharmacy or medical institution data. All methods performed better on the online pharmacy subset of the test bed (as compared to medical institutions). This is likely due to the small quantities of fraudulent medical institution instances in the dataset. RTL's AUC values were higher than the comparison methods, for both the balanced and imbalanced settings. The results further demonstrate the efficacy of RTL over existing adaptive methods.

6.6. Analysis of Performance on Different Medical Web Site Categories

In order to demonstrate RTL's effectiveness for detection of different categories of fake medical Web sites, we analyzed its performance on all three subsets of the test bed: online pharmacies, health information, and medical institution Web sites. The performance on each category was evaluated using overall accuracy and class-level f-measures, precision, and recall, averaged across the 30 bootstrap runs (as done in the previous experiments). Hence, for each of the 30 bootstrap runs, these performance metrics were computed on the three subsets of the 850 test instances. RTL's performance was analyzed in comparison with the best content, graph, stack, and adaptive methods reported in Sections 6.1–6.5.

Table XI shows the results. For each of the 5 methods (i.e., RTL, AZProtect, QoC, Logistic Regression stack, and SGL), the results on the each Web site category are depicted along with the overall results across all sites. In addition to being more accurate, RTL was also the most balanced. It attained overall accuracies in excess of 90% on all three Web site categories, outperforming the best comparison methods by at least 1.5%–6.5% on each category. The performance gains were largest on the health information Web sites.

Figure 16 shows a panoramic view of the 7 evaluation metrics on the overall test bed as well as the three categories. The figure hence depicts a sequential plot/curve

Table XI. Results by Web Site Category for RTL and Best Comparison Methods

Algorithm		Overall Accuracy	Legit			Fake		
			F-meas.	Prec.	Rec.	F-meas.	Prec.	Rec.
RTL	Pharmacy	97.84	96.08	93.01	99.37	98.51	99.77	97.28
	Health Info	91.25	90.86	97.72	84.90	91.61	86.06	97.92
	Medical Inst.	93.82	96.34	96.77	95.90	80.27	78.35	82.27
	Overall	94.33	93.83	96.22	91.55	94.76	92.80	96.80
SGL	Pharmacy	95.54	91.82	89.98	93.74	96.94	97.68	96.20
	Health Info	86.68	85.66	95.53	77.65	87.57	80.38	96.18
	Medical Inst.	88.75	93.20	95.52	90.98	67.47	60.43	76.36
	Overall	90.59	89.52	94.08	85.38	91.46	87.99	95.22
LRStk	Pharmacy	95.72	92.19	89.79	94.73	97.05	98.04	96.08
	Health Info	84.99	83.48	95.65	74.06	86.24	77.97	96.46
	Medical Inst.	90.14	94.01	96.79	91.39	72.05	63.54	83.18
	Overall	90.16	88.94	94.44	84.05	91.14	87.09	95.60
AZPro	Pharmacy	93.78	88.69	86.13	91.42	95.71	96.80	94.64
	Health Info	83.89	82.67	92.12	74.97	84.95	78.01	93.26
	Medical Inst.	88.06	92.71	95.96	89.67	66.92	58.00	79.09
	Overall	88.56	87.26	91.73	83.20	89.63	86.21	93.33
QoC	Pharmacy	94.82	90.23	91.45	89.27	96.47	96.14	96.84
	Health Info	79.28	77.25	88.29	68.66	80.98	73.31	90.43
	Medical Inst.	84.42	90.11	94.69	86.67	59.46	52.94	71.97
	Overall	86.39	84.50	91.02	78.85	87.86	83.19	92.38

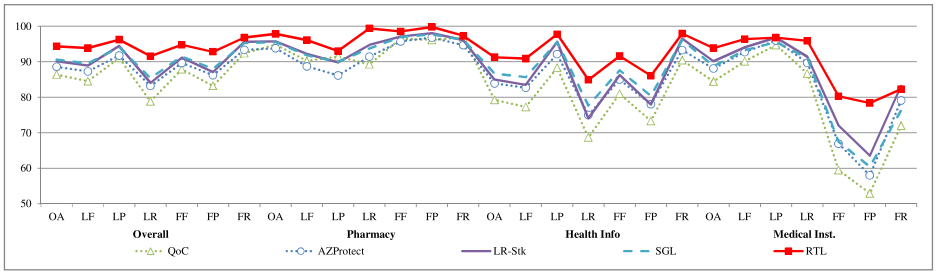

Fig. 16. Comparison of RTL and comparison methods on overall test bed and online pharmacy, health information, and medical institution Web site subsets.

of the 28 evaluation metrics associated with each of the four comparison methods. For instance, on the horizontal axis above "overall," LF, LP, and LR correspond to the overall legit f-measure precision, and recall. The figure further reaffirms RTL's effectiveness over the best comparison methods; it performed better on nearly every evaluation metric across Web site categories. These results suggest that RTL is better suited for detecting fake online pharmacy, health information, and medical institution Web sites.

7. RESULTS DISCUSSION

Based on the results presented in Section 6, RTL's enhanced performance was attributable to its content and graph classifiers, as well as the recursive labeling mechanism. Detailed analysis was performed in order to assess the impact of these component's parameter settings on RTL's overall performance. For all three components of

RTL, 9 different parameter combinations were run while holding the other two RTL component's parameter values constant (these were held at the values used in the experiments described in Section 5). The RTL-CC content classifier was run using values of 50, 100, and 200 for x (i.e., pages per Web site), and 10, 15, and 20 for l. The RTL-GC graph classifier was run using values of 0.3, 0.5, and 0.7 for α, and 0.25, 0.50, and 0.75 for β. The recursive labeling mechanism was run using $p = 5$ through $p = 200$, in various increments. The first three charts depicted in Figure 17 show the results for each component's 9 parameter combinations (top left/right and bottom left charts). For each chart, the x-axis displays the seven evaluation metrics. For example, OA denotes overall accuracy while LP and FR are legit precision and fake Web site recall, respectively.

The results from these three charts reveal that while the class-level f-measure, precision, and recall values fluctuated somewhat for different parameter settings, the overall accuracy values were fairly stable (ranging between 93.0% and 95.5%). Moreover, the parameter settings used in the experiments did not yield the best results. Running the graph classifier using $\alpha = 0.70$ and $\beta = 0.75$, while keeping the content classifier and recursive labeling mechanism's settings unchanged, improved overall accuracy to 95.28%. Similarly, running the content classifier using $x = 200$ improved overall accuracy to 95.54% (the setting used in the experiments was $x = 100$). As expected, the recursive labeling mechanism worked best when using smaller values for p (i.e., $p = 5$ through $p = 25$). However, the recursive labeling mechanism's parameter settings were less influential than the content and graph classifiers' in terms of their impact on performance. For instance, using $p = 5$ only yielded a 0.3% gain in accuracy over using $p = 50$.

In order to assess the combined effect of the three components' parameters on RTL's performance and computation time, we ran all possible combinations of the aforementioned parameter's values without holding any component's settings constant. This resulted in 729 parameter combinations (i.e., 9 x 9 x 9). For each setting, the bottom right chart in Figure 17 depicts the ratio of true positives to false positives on the y-axis (as a measure of overall effectiveness of each particular parameter combination), and RTL's average computation time per Web site, in seconds. Based on the chart, it is apparent that for most parameter settings, RTL's average computation time per Web site was less than 2 seconds. The recursive labeling component's p parameter had the greatest impact on computation times, with smaller values resulting in greater run times. For instance, the right-most cluster in the chart (at 4 seconds per Web site) are the results when p = 5, while the next cluster (at 2 seconds) are the results when p = 10. As previously alluded to, these higher run times did not lead to significant performance gains. The results reported in Section 6 (which used p = 50) had an average run time of 0.75 seconds per Web site. Overall, the analysis results suggest that RTL is fairly consistent across parameter settings with respect to its fake medical Web site detection capabilities.

7.1. Analysis of RTL's Recursive Labeling Mechanism's Performance

We assessed the impact of alternate stopping rules and the performance of RTL-GC and RTL-CC on the performance of the recursive labeling mechanism. For the stopping rule analysis, we compared the current rule, where all testing instances were eventually added to the training model (Add-All), against a rule where the recursive labeling stopped when the remaining additional instances had a rank of $2m$ (Stop-$2m$) and a rule where all the $2m$ instances were ranked in descending order (Rank-$2m$) based on the difference between the absolute values of the prediction scores associated with RTL-CC and RTL-GC (i.e., $CC(S) + GC(S)$). The major difference between

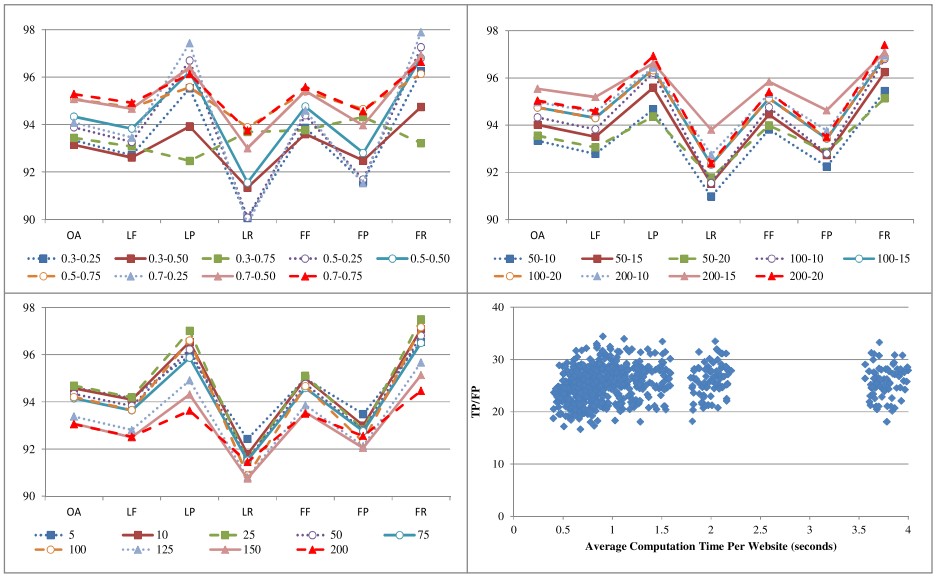

Fig. 17. Impact of graph classifier (top left), content classifier (top right), recursive labeling mechanism (bottom left), and combinations of parameter settings (bottom right) on RTL's performance.

Table XII. Results Using Different Stopping Conditions for RTL's Recursive Labeling Mechanism

Algorithm	Overall Accuracy	Legit			Fake		
		F-meas.	Prec.	Recall	F-meas.	Prec.	Recall
Stop-2m	94.64	94.15	96.68	91.75	95.05	92.99	97.20
Rank-2m	−0.12	−0.13	−0.16	−0.10	−0.11	−0.09	−0.13
Add-All	−0.31	−0.33	−0.46	−0.20	−0.29	−0.19	−0.40

Add-All and Rank-2m was that for the former, 2m instances were added arbitrarily to the training models, while for the latter, instances where one of the underlying classifiers was more confident about its predictions were added earlier. Table XII shows the evaluation results. Both the Stop-2m and Rank-2m stopping conditions attained better results than the Add-All approach adopted in this study. The results suggest that filtering at least some of the 2m instances (i.e., ones where RTL-CC and RTL-GC disagree) during the recursive labeling phase may be beneficial. Additional analysis revealed that the number of 2m instances decreased across iterations of the recursive labeling mechanism, therefore, the number of 2m instances actually added to the training data signified a small percentage. Consequently, the performance differences for the three stopping conditions were not major. However, the performance of these different stopping conditions at an iteration-by-iteration level revealed that some 2m instances did increase discriminatory potential when incorporated in the training set. Based on these findings, future work that explores the impact of more in-depth stopping conditions that can better leverage such test instances may be warranted.

We also conducted analysis in order to gain empirical insights regarding the impact of the underlying RTL-CC and RTL-GC classifiers' performances on the effectiveness of the recursive labeling mechanism. In our analysis of RTL across the 30 bootstrap iterations, we observed that the performance of the recursive labeling mechanism seemed to be somewhat impacted by the average accuracy across the two underlying classifiers (RTL-CC and RTL-GC), as well as the difference in accuracies between these

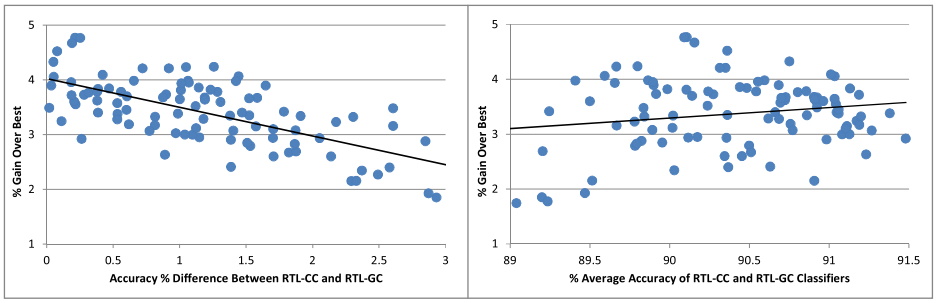

Fig. 18. Impact of RTL-CC and RTL-GC performance on effectiveness of recursive labeling mechanism.

two classifiers. To further explore these relations, we ran 100 bootstrap iterations of RTL. Figure 18 shows the results. Each point represents one of the 100 runs. Both charts depict the percentage gain in accuracy when using recursive labeling (on the y-axis) over whichever underlying classifier had better performance. For the chart on the right, the x-axis displays the average percentage accuracy of the two underlying classifiers. The chart on the left shows the percentage difference in accuracy between the two underlying classifiers on the x-axis. Based on the trend lines, we note that the recursive labeling mechanism's margin of improvement over the best underlying classifier was greater when the performance difference between RTL-CC and RTL-GC was lower, and to a lesser extent when the average accuracy of RTL-CC and RTL-GC was higher. In other words, the underlying classifiers complemented each other the most during the recursive labeling phase when their accuracies were comparable and higher.

8. CONCLUSIONS

In this work we proposed RTL, a metalearning algorithm that uses an adaptive learning mechanism in conjunction with information from underlying content and graph-based classifiers specifically designed to exploit the characteristics of online medical content. RTL-CC and RTL-GC were able to significantly improve results over various comparison graph and content classifiers. RTL attained an overall accuracy of over 94% and outperformed comparison stack classifiers and adaptive methods, demonstrating its viability for detection of fake medical Web sites. Additionally, robustness analysis results revealed that it was less susceptible to poor results when dealing with limited quantities of training Web sites. RTL was generally able to perform well when using as little as 15–30 training Web sites. Analysis of the performance results across medical Web site categories showed that RTL was fairly balanced in its detection capabilities; it attained over 90% accuracy on all three test bed subsets (online pharmacy, health information, and medical institution Web sites). Further detailed analysis demonstrated RTL's effectiveness across different parameter settings; it was able to consistently and efficiently distinguish legitimate medical Web sites from fake ones at a high level of accuracy.

The experimental results provide several important insights. RTL-GC's use of simultaneous dual class propagation across in/out links was more effective than existing graph-based methods. The complexity of content associated with the medical domain necessitates the use of more robust, domain-specific content classifiers. RTL-CC's use of a medical thesaurus and redundancy reducing feature extractor allowed a better coverage of medical concepts, and played an integral role in RTL's enhanced detection of fake health information sites. RTL's combination of robust underlying classifiers

and instance-centric recursive labeling mechanism made it more effective than stack classifiers and existing adaptive methods which rely on iterative feature construction.

In our future work, we intend to further assess the impact of different stopping rules and instance ranking strategies on the performance of the recursive labeling mechanism. We also intend to examine the potential to develop additional empirical insights about RTL's likelihood to improve over its underlying classifiers. Given the dire social implications associated with the rampant sale of fake pharmaceuticals, phony medical institutions, and increased medical misinformation, the results have important implications for online trust and security in the era of Health 2.0.

REFERENCES

ABBASI, A. AND CHEN, H. 2009a. A comparison of tools for detecting fake websites. *IEEE Comput. 42*, 10, 78–86.

ABBASI, A. AND CHEN, H. 2009b. Comparison of fraud cues and classification methods for fake escrow website detection. *Inf. Techn. Manage. 10*, 2, 838–101.

ABBASI, A., ZHANG, Z., ZIMBRA, D., CHEN, H., AND NUNAMAKER JR., J. F. 2010. Detecting fake websites: The contribution of statistical learning theory. *MIS Quart. 34*, 3, 435–461.

ABERNATHY, J., CHAPELLE, O., AND CASTILLO, C. 2008. Web spam identification through content and hyperlinks. In *Proceedings of the 4th International Workshop on Adversarial Information Retrieval on the Web*. 41–44.

ADOMAVICIUS, G. AND TUZHILIN, A. 2005. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Trans. Knowl. Data Engin. 17*, 6, 734–749.

ALDHOUS, P. 2005. Counterfeit pharmaceuticals: Murder by medicine. *Nature 434*, 132–136.

AN, B. 2010. 14 Arrested for making, selling fake drugs via bogus military medical websites. *Xinhua Net* (2/2/10). http://big5.xinhuanet.com/gate/big5/news.xinhuanet.com/english2010/china/2010-02/05/c_13165317.htm.

ANDERSON, B. AND MOORE, A. 1998. ADtrees for fast counting and for fast learning of association rules. In *Proceedings of the Conference on Knowledge Discovery from Databases*. 134–138.

APHINYANAPHONGS, Y. AND ALIFERIS, C. F. 2007. Text categorization models for identifying unproven cancer treatments on the web. In *Proceedings of the World Congress on Health (Medical) Informatics: Building Sustainable Health Systems*. 968–972.

ARAUJO, L. AND MARTINEZ-ROMO, J. 2010. Web spam detection: New classification features based on qualified link analysis and language models. *IEEE Trans. Inf. Forensics Secur. 5*, 3, 581–590.

ARMIN, J. 2010. Internet drug rings and their 'killer' online pharmacies. *Internet Evolution* (5/10/10). http://www.internetevolution.com/author.asp?section_id=717&doc_id=191640.

BAYES, T. 1958. Studies in the history of probability and statistics: XI. Thomas Bayes' essay towards solving a problem in the doctrine of chances. *Biometrika 45*, 293–295.

BECCHETTI, L., CASTILLO, C., DONATO, D., LEONARDI, S., AND BAEZA-YATES, R. 2006. Link-based characterization and detection of web spam. In *Proceedings of the 2nd International Workshop on Adversarial Information Retrieval on the Web*.

BECCHETTI, L., CASTILLO, C., DONATO, D., BAEZA-YATES, R., AND LEONARDI, S. 2008. Link analysis for web spam detection. *ACM Trans. Web 2*, 1, Article 2.

BOGGAN, S. 2009. Headache pills made of rat poison... *The Daily Mail* (4/27/09). http://www.thefreelibrary.com/Headache+pills+made+of+rat+poison.+Viagra+made+of+chalk.+After+a+...-a0198806854.

BRAZDIL, P., GIRAUD-CARRIER, C., SOARES, C., AND VILALTA, R. 2008. *Metalearning: Applications to Data Mining*. Springer-Verlag, Berlin.

CASTILLO, C., DONATO, D., GIONIS, A., MURDOCK, V., AND SILVESTRI, F. 2007. Know your neighbors: Web spam detection using the web topology. In *Proceedings of the ACM Special Interest Group on Information Retrieval*. 423–430.

CHEN, H., LALLY, A. M., ZHU, B., AND CHAU, M. 2003. HelpfulMed: Intelligent searching for medical information over the internet. *J. Amer. Soc. Inf. Sci. Techn. 54*, 7, 683–694.

CHOU, N., LEDESMA, R., TERAGUCHI, Y., BONEH, D., AND MITCHELL, J. C. 2004. Client-side defense against web-based identity theft. In *Proceedings of the Network and Distributed System Security Symposium*.

CHUA, C. E. H. AND WAREHAM, J. 2004. Fighting internet auction fraud: An assessment and proposal. *IEEE Comput. 37*, 10, 31–37.

CHUA, C. E. H., WAREHAM, J., AND ROBEY, D. 2007. The role of online trading communities in managing internet auction fraud. *MIS Quart. 31*, 4, 759–781.

DINEV, T. 2006. Why spoofing is serious internet fraud. *Comm. ACM 49*, 10, 76–82.

DROST, I. AND SCHEFFER, T. 2005. Thwarting the nigritude ultramarine: Learning to identify link spam. In *Proceedings of the European Conference on Machine Learning*. 96–107.

EASTON, G. 2007. Clicking for pills. *Brit. Med. J. 334*, 7583, 14–15.

ESTER, M., KRIEGEL, H., AND SCHUBERT, M. 2002. Web site mining: A new way to spot competitors, customers, and suppliers in the World Wide Web. In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 249–258.

EYSENBACH, G. 2008. Medicine 2.0: Social networking, collaboration, participation, apomediation, and openness. *J. Med. Intern. Resear. 10*, 3, e23.

FANG, X., SHENG, O. R. L., AND CHAU, M. 2007. ServiceFinder: A method towards enhancing service portals. *ACM Trans. Inf. Syst. 25*, 4, Article 17.

FETTERLY, D., MANASSE, M., AND NAJORK, M. 2004. Spam, damn spam, and statistics. In *Proceedings of the 7th International Workshop on the Web and Databases*. 1–6.

FU, A. Y., LIU, W., AND DENG, X. 2006. Detecting phishing web pages with visual similarity assessment based on earth movers distance (EMD). *IEEE Trans. Depend. Secure Comput. 3*, 4, 301–311.

GAN, Q. AND SUEL, T. 2007. Improving web spam classifiers using link structure. In *Proceedings of the 3rd International Workshop on Adversarial Information Retrieval on the Web*. 17–20.

GAUDINAT, A., GRABAR, N., AND BOYER, C. 2007. Machine learning approach for automatic quality criteria detection of health web pages. In *Proceedings of the World Congress on Health (Medical) Informatics: Building Sustainable Health Systems*. 705–729.

GRAZIOLI, S. AND JARVENPAA, S. L. 2000. Perils of internet fraud: An empirical investigation of deception and trust with experienced internet consumers. *IEEE Trans. Syst. Man. Cyber. Part A 20*, 4, 395–410.

GREENBERG, A. 2008. Pharma's black market boom. *Forbes.com* (8/26/08).

GYONGYI, Z. AND GARCIA-MOLINA, H. 2005. Spam: It's not just for inboxes anymore. *IEEE Comput. 38*, 10, 28–34.

GYONGYI, Z., GARCIA-MOLINA, H., AND PEDERSEN, J. 2004. Combating web spam with trust rank. In *Proceedings of the 13th International Conference on Very Large Data Bases*. 576–587.

GYONGYI, Z. BERKHIN, P., GARCIA-MOLINA, H., AND PEDERSEN, J. 2006. Link Spam detection based on mass estimation. In *Proceedings of the ACM Conference on Very Large Databases*. 439–450.

HESSE, B. W., HANSEN, D., FINHOLT, T., MUNSON, S., KELLOGG, W., AND THOMAS, J. C. 2010. Social participation in health 2.0. *IEEE Comput. 43*, 11, 45–52.

HUGHES, B., JOSHI, I., AND WAREHAM, J. 2008. Health 2.0 and Medicine 2.0: Tensions and controversies in the field. *J. Med. Intern. Resear. 10*, 3, e23.

JOACHIMS, T. 2002. *Learning to Classify Text Using Support Vector Machines: Methods, Theory, and Algorithms*. Kluwer Academic Publishers.

KOLARI, P., FININ, T., AND JOSHI, A. 2006. SVMs for the blogosphere: Blog identification and splog detection. In *Proceedings of the AAAI Spring Symposium on Computational Approaches to Analysing Weblogs*.

KREBS, B. 2005. Few online 'Canadian pharmacies' based in Canada, FDA says. *WashingtonPost.com* (6/14/05).

KRIEGEL, H. AND SCHUBERT, M. 2004. Classification of websites as sets of feature vectors. In *Proceedings of the International Conference on Databases and Applications*. 127–132.

KRISHNAN, V. AND RAJ, R. 2006. Web spam detection with anti-trust rank. In *Proceedings of the 2nd International Workshop on Adversarial Information Retrieval on the Web*. 37–40.

LE, A., MARKOPOULOU, A., AND FALOUTSOS, M. 2011. PhishDef: URL Names say it all. In *Proceedings of the IEEE International Conference on Computer Communications*.

LIN, F. C., SHI, H., AND WANG, X. 2007. Splog detection using self-similarity analysis on blog temporal dynamics. In *Proceedings of the 3rd International Workshop on Adversarial Information Retrieval on the Web*.

LIU, H., JOHNSON, S. B., AND FRIEDMAN, C. 2002. Automatic resolution of ambiguous terms based on machine learning. *J. Amer. Med. Informatics Assoc. 9*, 6, 621–636.

LIU, W., DENG, X., HUANG, G., AND FU, A. Y. 2006. An antiphishing strategy based on visual similarity assessment. *IEEE Intern. Comput. 10*, 2, 58–65.

LUO, G. 2008. MedSearch: A specialized search engine for medical information retrieval. In *Proceedings of the 17th ACM Conference on Information Knowledge and Management*. 143–152.

MARTINEZ-ROMO, J. AND ARAUJO, L. 2009. Web spam identification through language model analysis. In *Proceedings of the 5th International Workshop on Adversarial Information Retrieval on the Web*. 21–28.

NIE, L., WU, B., AND DAVISON, B. D. 2007a. Winnowing wheat from the chaff: Propagating trust to sift spam from the web. In *Proceedings of the Workshop of the ACM Special Interest Group on Information Retrieval*. 869–870.

NIE, L., WU, B., AND DAVISON, B. D. 2007b. A cautious surfer for PageRank. In *Proceedings of the 17th International World Wide Web Conference*. 1119–1120.

NTOULAS, A., NAJORK, M., MANASSE, M., AND FETTERLY, D. 2006. Detecting spam web pages through content analysis. In *Proceedings of the 15th International World Wide Web Conference*. 83–92.

QUINLAN, R. 1986. Induction of decision trees. *Mach. Learning 1*, 1, 181–106.

PAGE, B., BRIN, S., MOTWANI, R., AND WINOGRAD, T. 1998. The PageRank citation ranking: Bringing order to the web. Tech. rep., Stanford University.

PARLOFF, R. 2010. New legal trick: Fake hospital sites for finding clients. *CNN* (4/1/10) http://money.cnn.com/2010/04/01/news/companies/fake_va_hospitals_websites.fortune/.

PEW INTERNET AND AMERICAN LIFE PROJECT. 2009. http://www.pewinternet.org/Press-Releases/2009/The-Social-Life-of-Health-Information.aspx.

PRICE, S. L. AND HERSH, W. R. 1999. Filtering web pages for quality indicators: An empirical approach to finding high quality consumer health information on the world wide web. In *Proceedings of the AMIA Symposium*. 911–915.

RILOFF, E., PATWARDHAN, S., AND WIEBE, J. 2006. Feature subsumption for opinion analysis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. 440–448.

SALVETTI, F. AND NICOLOV, N. 2006. Weblog classification for fast splog filtering: A URL language model segmentation approach. In *Proceedings of the Human Language Technology Conference*. 137–140.

SHEN, G., GAO, B., LIU, T. Y., FENG, G., SONG, S., AND LI, H. 2006. Detecting link spam using temporal information. In *Proceedings of the International Conference on Data Mining*.

SONDHI, P., VYDISWARAN, V. G. V., AND ZHAI, C. 2012. Reliability prediction of webpages in the medical domain. In *Proceedings of the 34th European Conference on Information Retrieval*. 219–231.

SONG, J. AND ZAHEDI, F. M. 2007. Trust in Health Infomediaries. *Decision Support Syst. 43*, 390–407.

TIAN, Y., WEISS, G. M., AND MA, Q. 2007. A semi-supervised approach for web spam detection using combinatorial feature-fusion. In *Proceedings of the ECML Graph Labeling Workshops' Web Spam Challenge*.

URVOY, T., LAVERGNE, T., AND FILOCHE, P. 2006. Tracking web spam with hidden style similarity. In *Proceedings of the 2nd International Workshop on Adversarial Information Retrieval on the Web*.

URVOY, T., CHAVEAU, E., FILOCHE, P. AND LAVERGNE, T. 2008. Tracking web spam with hidden style similarities. *ACM Trans. Web 2*, 1, Article 3.

WANG, X., TAO, T., SUN, J., SHAKERY, A., AND ZHAI, C. 2008. DirichletRank: Solving the zero-one gap problem of PageRank. *ACM Trans. Inf. Syst. 26*, 2, Article 10.

WANG, Y. AND RICHARD, R. 2007. Rule-based automatic criteria detection for assessing quality of online health information. *J. Inf. Techn. Healthcare 5*, 5, 288–299.

WHITE, R. W. AND HORVITZ, E. 2009. CyberChondria: Studies of the escalation of medical concerns in web search. *ACM Trans. Inf. Syst. 27*, 4, Article 23.

WILFORD, B. B., SMITH, D. E., AND BUCHER, R. 2006. Prescription stimulant sales on the internet. *Pediatric Annals 35*, 8, 575–578.

WITTEN, I. H. AND FRANK, E. 2005. *Data Mining: Practical Machine Learning Tools and Techniques* 2nd Ed. Morgan Kaufmann, San Francisco, CA.

WU, B. AND CHELLAPILLA, K. 2007. Extracting link spam using biased random walks from spam seed sets. In *Proceedings of the 3rd International Workshop on Adversarial Information Retrieval on the Web*. 37–44.

WU, B. AND DAVISON, B. D. 2005. Identifying link farm spam pages. In *Proceedings of the 14th International World Wide Web Conference*. 820–829.

WU, B. AND DAVISON, B. D. 2006. Detecting semantic cloaking on the web. In *Proceedings of the 15th International World Wide Web Conference*. 819–828.

WU, B., GOEL, V., AND DAVISON, B. D. 2006. Propagating trust and distrust to demote web spam. In *Proceedings of the Workshop on Models of Trust for the Web*.

ZAHEDI, F. M. AND SONG, J. 2008. Dynamics of trust revision: Using health infomediaries. *J. Manage. Inf. Syst. 24*, 4, 225–248.

ZHANG, L., ZHANG, Y., AND ZHANG, Y., AND LI, X. 2006. Exploring both content and link quality for anti-spamming. In *Proceedings of the 6th IEEE International Conference on Computer and Information Technology*. 37–42.

ZHOU, B. AND PEI, J. 2009. Link spam target detection using page farms. *ACM Trans. Knowl. Discov. Data 3*, 3, Article 13.

Received December 2010; revised August 2011, January 2012; accepted March 2012