Спам в соц. сетях передается в личных сообщениях, показывается в новостных лентах, на стенах поддельных аккаунтов, принадлежащих разным типам спам-ботов (displayers, braggers, posters, whisperers). Сложнее всего детектировать аккаунты displayer'ов и bragger'ов. Они не производят непосредственной рассылки сообщений или их публикации на стенах "друзей-жертв", а демонстрируют всю рекламу в своих лентах и альбомах, добавляясь в друзья, подписчики к легитимным, чаще всего новым пользователям.

Почему с аккаунтами подобных спам-ботов надо бороться:

- аккаунты displayer'ов и bragger'ов в первую очередь используются для распространения URL phishing-сайтов, сайтов с зараженным контентом, во вторую очередь - для распространения рекламы;
- паразитные аккаунты генерируют паразитный трафик, а их контент требует штатного обслуживания: бекапирования, сохранения целостности, индексации, он занимает место в хранилищах данных, мешает при проведении статистических оценок и исследований;
- аккаунты displayer'ов и bragger'ов отталкивают новых пользователей OSN навязчивой рекламой, провокативными репостами, потенциальной угрозой безопасности.

Все это формирует плохую репутацию OSN в СМИ и среди людей. Стоимость регистрации новых аккаунтов в соц. сетях в связи с развитием технологий падает, интернет-маркетинг и использование электронных платежей среди рядовых пользователей Рунета постоянно растет => подобные "неагрессивные" (с т.з. рассылки сообщений и "Posting behaviour") аккаунты displayer'ов и bragger'ов будут оставаться выгодными для phisher'ов, carder'ов, владельцев ботнетов.

Согласно закону Парето, около 20% легитимных пользователей OSN постоянно генерирует свой оригинальный, интересный контент, позволяющий привлекать новых пользователей в соц. сеть - первый тип аккаунтов. Этот тип легко детектируются благодаря "сильным", общеизвестным метрикам и шаблонам. Остальная часть валидных аккаунтов (80%) также периодически добавляет оригинальный, новый контент, но в подавляющем числе случаев делает репосты, оставляет ссылки на уже опубликованные тексты, картинки, видео и пр. - второй тип аккаунтов. Естественно, цель displayer'ов и bragger'ов - делать свои страницы максимально похожими на второй тип. Для него также существуют свои шаблоны и "сильные" метрики. Но надо отметить, что некоторые из распространенных метрик в случае displayer'ов и bragger'ов никак не влияют на качество детектирования их страниц в OSN, а в определенных условиях могут даже спровоцировать повышенный false-posifive rate:

1. Sender Social Degree: сильный признак, основанный на количестве связей аккаунта с другими в соц. графе.

Принятно считать, что для поддельных аккаунтов данная метрика традиционно высокая. Но этот признак имеет высокое значение и для легитимных пользователей, активно генерирующих новый контент, для известных людей, организаций, агентов по поиску новых клиентов и т.п.

У аккаунтов displayer'ов и bragger'ов большое число связей как с легитимными пользователями, так и друг с другом, они часто могут находиться в центрах кластеров валидных пользователей на социальном графе.

2. Posting Behaviour Pattern: шаблон, базовые метрики которого: частота и регулярность новых публикаций, средняя длина публикуемых текстов, timestamp, наличие URL, графики, количество графики и т.п.

Для displayer'ов и bragger'ов данный шаблон по значениям метрик вполне схож с аналогичным шаблоном для второго типа аккаунтов: новые сообщения/репосты на стенах появляются примерно в определенное время, с регулярностью - 2-3 раза в день, на 4-5 "чистых" сообщений в ленте – 1-2 поста с ссылками, ведущими на зараженный или phising-сайт.

3. Account LifeTime: аккаунты displayer'ов и bragger'ов существуют достаточно приличное время с момента регистрации, как и аккаунты первого типа, некоторые из аккаунтов второго типа.

4. Average Num URL per message: для легитимных пользователей этот показатель также достаточно высок.

5. Social Behaviour Pattern: шаблон, базовые метрики которого: периодичность и отношение количество/время для friend-follower-запросов, частота и регулярность участия в голосованиях ("лайки"), пр. активность аккаунта.

Для аккаунтов displayer'ов и bragger'ов данный шаблон по значениям базовых метрик практически не отличается от того, что ожидается для пользователей второго типа. Аккаунты displayer'ов и bragger'ов периодически отправляют friend/follower-запросы либо к таким же поддельным аккаунтам, либо к новым пользователям, расставляют "лайки" на сообщения любых пользователей в лентах.

6. метрики, основанные на Bursty Behaviour Pattern: displayer'ы и bragger'ы осуществляют свою активность по методу Sybil-атак, поэтому прибыль от них напрямую определяется количеством их аккаунтов в OSN. Для легитимных пользователей "bursty behaviour" наоборот иногда бывает очень типичным в случае каких-нибудь экстраординарных общественных событий или праздников.

7. Контентный анализ сообщений, публикуемых поддельными аккаунтами показывает, что примерно 70% из них содержат ссылки на безопасные ресурсы, безобидный текст, графику. Этот контент позволяет displayer'ам и bragger'ам "обелять" репутацию своих аккаунтов и избегать блокировок.

Суть идеи – улучшить существующие антиспам- и репутационные фильтры, чтобы эффективно детектировать сообщества displayer'ов, bragger'ов внутри кластеров легитимных пользоваталей, быстро адаптироваться к появлению новых типов поддельных аккаунтов, к их шаблонам поведения и к новым видам спам-сообщений.

Способы реализации и обобщенная схема
(кратко перечислены основные шаги)

1. Displayer'ы и bragger'ы генерируют сообщения автоматически по заданным шаблонам, поэтому их посты легко кластеризовать с использованием алгоритмов k-means или c-means. Это позволит "на лету" выделять актуальные наборы метрик, которыми характеризуются текущие шаблоны ботов. Эти наборы метрик можно расширить стандартными признаками для спам-сообщений. На основе расширенного набора признаков можно построить классификатор сообщений (SVM или Decision Tree).

Классификатор будет выносить вердикт по каждому новому посту из ленты. И, таким образом, любому аккаунту, который обновляет свою ленту новостей, можно поставить в соответствие отношение числа spam-постов к ham-постам, сгенерированных им за время жизни. Данную метрику обычно принимают

за репутацию аккаунта. Аккаунты, ни разу не генерившие сообщений в ленте, должны иметь среднее значение репутации.

Каждые $N_{threshold}$ новых сообщений должны инкрементированно обновлять результаты кластеризации, т.е. они либо формируют свой новый кластер постов, либо добавляются в уже существующий, либо вызывают слияние существующих и т.д.

Кластеры сообщений также должны иметь коэффициент "затухания". Он будет зависеть от количества сообщений внутри кластера и от выбранного интервала времени: т.е. кластер разрушается, если внутри него число постов меньше порогового значения. Старые посты внутри кластера регулярно детектируются по timestamp и удаляются => размер кластера повторно переопределяется и сравнивается с порогом.

2. На каждое пользовательское событие в рамках одного аккаунта (событие - аккаунт создал новую связь - friend/follower request, проголосовал за какой-либо объект, загрузил медиа-контент на свою страницу и т.п.) формировать векторное представление данного аккаунта на основе шаблона или нескольких шаблонов, наиболее подходящих к типу случившегося события (Posting Behaviour Pattern, Social Behaviour Pattern, Hosting Behaviour Pattern, Interectaion History Pattern).

Далее расширить векторное представление аккаунта текущим значением его репутации и передать это векторное представление на анализ классификатору из семейства RandomForest. RandomForest должен вернуть для аккаунта значение класса: "валидный" или "поддельный".

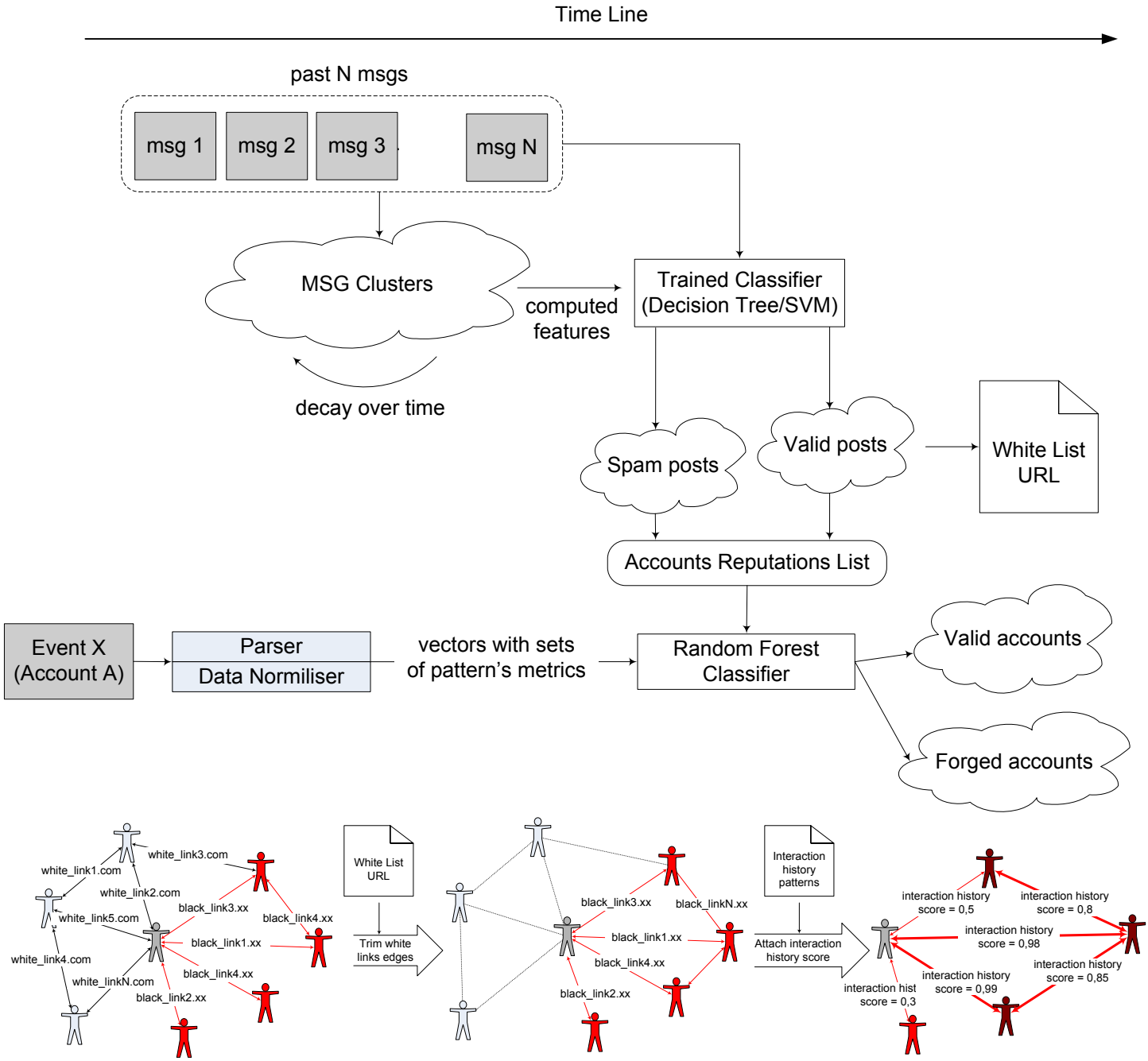
3. Если аккаунт валидный - сформировать список URL, которые могут присутствовать в его сообщениях, найденных в текущих кластерах постов, а затем смержить этот список ссылок с общим White URL List.

4. Построить социальный граф, где для ребер, представляющих связи узла с другими аккаунтами, в качестве весов использовать последнюю ссылку, которую пользователь публиковал в своей ленте.

Если для обрабатываемого аккаунта в кластерах сообщений не было найдено никаких ссылок => не указывать на графе его связи с другими узлами, т.е. вершина аккаунта будет изолированной. Далее убрать дуги, вес которых определяют ссылки из White URL List => получим еще некоторый набор изолированных вершин. Необходимо исключить все изолированные вершины. Таким образом, в полученном подграфе останутся подозрительные узлы. Среди этих подозрительных узлов, возможно, будут те, которые принадлежат displayer'ам и bragger'ам.

5. Чтобы еще больше сузить круг поиска аккаунтов displayer'ов и bragger'ов в полученном подграфе, нужно обратиться к истории взаимодействия пользователей друг с другом. Пользователь может устанавливать большое число связей с другими пользователями, но для любого легитимного аккаунта регулярный и частый обмен сообщениями, постами/репостами, "лайками", подарками, приложениями происходит в среднем только с 5-20 "друзьями". Соотвественно, можно сформировать такую метрику как interaction history score. Если любой тип взаимодействия между парой связанных аккаунтов обозначить как K[i], то interaction history score = 1/sum(K[i]). И тогда на полученном подграфе дуги взаимодействий между узлами, которые редко "общаются", будут иметь больший interaction history score => будут соотвествовать дугам, связывающим узлы displayer'ов и bragger'ов.

Displayer'ы и bragger'ы, чтобы оставаться незаметными, только публикуют спам в своих лентах, редко шлют личные сообщения "друзьям", ничего им не дарят, не пользуются приложениями, не публикуют видео или музыку в своих альбомах, т.е. их связи с легитимными аккаунтами и друг с другом всегда будут иметь высокий interaction history score.



Общие идеи по улучшению детекта поддельных аккаунтов существующими антиспам- и репутационными фильтрами:

1. формирование Posting Behaviour Pattern, Social Behaviour Pattern для аккаунтов displayer'ов и bragger'ов с раширенным набором метрик:

- метрики на основе числа и типов отправляемых подарков, используемых приложений, просматриваемого и добавляемого медиа-контента (видео, музыка);
- другие метрики, не коррелирующие с временными интервалами и количественными характеристиками, т.к. подобные признаки легко обходятся в случае Sybil Attacks;
- дельта между временем регистрации аккаунта и первым постом; метрики, основанные на "истории жизни" аккаунта.

2. более детальное накопление и изучение истории взаимодействия между узлами на соц. графе - формирование Interaction History Pattern с признаками для двух основных категорий пользователей + аккаунтов displayer'ов и bragger'ов.

3. использование метрики Unique URL Number, формирование шаблонов URL, распространяемых в ходе спам-кампаний;

Unique URL Number: снижение стоимости хостинга, наличие short-url сервисов, возможность заводить домены в интернационализированных зонах типа .рф позволяют спам-ботам генерировать большое число уникальных ссылок для обхода Black URL Lists, в то время как большинство валидных аккаунтов обычно постит от 2-х до 5-ти ссылок на одни и те же общеизвестные сайты или нашумевшие ресурсы. Этот факт также оправдывает использование глобального белого списка ссылок.

4. формирование метрик, описывающих Hosting Behaviour Pattern, на основе анализа распределения IP-адресов хостов, с которых производятся REST-запросы к API OSN.

Для displayer'ов и bragger'ов, отправляющих запросы и имитирующих активность очень большого числа поддельных аккаунтов, количество уникальных IP и доменов, откуда отправляются запросы, должно быть небольшим.

5. для задачи выделения "сильных" признаков, позволяющих классифицировать текстовые сообщения попробовать использовать комбинации таких feature selection-алгоритмов как Variance Threshhold + k-best, L1 + k-best. Графику, входящую в состав сообщений классифицировать отдельно с помощью NeuralNet и результат классификации "суммировать" с результатами для текстовых шинглов и ссылок.