УДК 519.16:621.395

Главацкий С.П.
Главацький С.П.
Glavatskiy S. P.

СТАТИСТИЧЕСКИЙ АНАЛИЗ ТРАФИКА СОЦИАЛЬНЫХ СЕТЕЙ

СТАТИСТИЧНИЙ АНАЛІЗ ТРАФІКА СОЦІАЛЬНИХ МЕРЕЖ

STATISTICAL ANALYSIS OF THE SOCIAL NETWORK TRAFFIC

Аннотация. Проведено исследование трафика наиболее популярных социальных сетей при помощи статистических методов анализа. В результате анализа получены данные о количестве активных пользователей и объеме трафика исследуемых сетей. Полученные данные позволяют оценить популярность различных социальных сетей и оценить их долю трафика, по сравнению с другими сервисами Интернета. Подобные исследования для Украинского сегмента сети интернет не производились, в результате чего сравнение производилось с Российскими исследованиями трафика социальных сетей. Результаты показывают различия используемых сервисов в зависимости от географического положения. Результатом исследования является доказательство самоподобных свойств трафика социальных сетей, рассчитаны показатели Херста для трафика в зависимости от дня недели и исследуемой сети.

Анотація. Проведено дослідження трафіка найбільш популярних соціальних мереж за допомогою статистичних методів аналізу. У результаті аналізу отримано дані про кількість активних користувачів і обсягу трафіка досліджуваних мереж. Отримані дані дозволяють оцінити популярність різних соціальних мереж і оцінити їх частку трафіка, порівнянно з іншими сервісами Інтернету. Подібні дослідження для Українського сегмента мережі Інтернет не проводилися, в результаті чого порівняння проводилося з Російськими дослідженнями трафіка соціальних мереж. Результати показують відмінності використовуваних сервісів залежно від географічного положення. Результатом дослідження є доказ самоподібних властивостей трафіка соціальних мереж, розраховані показники Херста для трафіка в залежності від дня тижня і досліджуваної мережі.

Summary. Research of a traffic of the most popular social networks by means of statistical methods of the analysis is conducted. As a result of the analysis data on number of active users and volume of a traffic of studied networks are obtained. The obtained data allow to estimate popularity of various social networks and to estimate their share of a traffic, in comparison with other services of the Internet. Similar researches weren't made for the Ukrainian segment of the Internet therefore comparison was made with the Russian researches of a traffic of social networks. Results show distinctions of used services depending on a geographical position. Result of research is the proof of self-similar properties of a traffic of social networks, Hurst exponents for a traffic depending on a day of the week and a studied network are calculated.

В последнее время в сети Интернет наблюдается тенденция увеличения количества пользователей различных социальных сетей. Сервис социальных сетей представляет собой сложную систему распределения трафика между тысячами серверов и миллионами пользователей. Последние исследования трафика компьютерных сетей доказывают, что он имеет особую фрактальную структуру и обладает свойством самоподобия, сохраняющуюся на различных временных промежутках. Самоподобные свойства трафика приводят к тому, что даже при небольшом объеме трафика могут наблюдаться резкие всплески трафика. Этот явление не описано в классической теории телетрафика с пуассоновским и марковским распределением и может существенно влиять на потери и механизмы качества обслуживания. При распределении трафика, обладающего эффектом самоподобия, использование классических алгоритмов распределения неэффективно и может привести к перегрузке одного из узлов. В связи с проявлением новых свойств трафика актуальны исследования его свойств для быстро развивающихся сервисов социальных сетей, для проектирования алгоритмов распределения самоподобного трафика.

Существенный вклад в изучение и анализ принципов организации социальных сетей внесли Г.В. Градосельская, Д. А. Губанова, Д. А. Новикова, А. Г. Чхартишвили, А.Н. Чураков [1, 2], однако их исследования носят более теоретических характер и не показывают степень использования Пользователями различных социальных сетей, тем более, что эти показатели меняются как в пространстве, так и во времени. Современные исследования, проведенные маркетинговым агентством “Редкая марка”[3] посредством среза данных из открытых источников, показывают значительную долю количества пользователей социальных сетей. Самоподобные свойства различных видов сетевого трафика были исследованы такими учеными, как: Б.С. Цыбаков, А.А. Потапов, Н.С. Лиханов [4, 5, 6],

однако исследование трафика социальных сетей с учетом эффекта самоподобия еще не проводилось, несмотря на то, что трафик социальных сетей является весомым и составляет существенную часть глобального Интернет трафика [3].

Целью настоящей статьи является исследование количественных и качественных характеристик наиболее популярных социальных сетей в Украине.

1. Описание методики исследования. На основании изученной литературы об устройстве сети Интернет [7,8] следует, что основным протоколом для сети Интернет есть протокол bgp, основным элементом которого есть так называемый номер автономной системы, который представляет собой подобие логина для авторизации, например AS35245. К номеру автономной системы прикрепляются сети ip-адресов. Все сервисы социальных сетей имеют свою независимую сеть ip-адресов. На основе публичных данных о подсетях, закрепленных за данным номером автономной системы, были определены все ip-адреса исследуемых социальных сетей, которые показаны в табл. 1.

Таблиця 1 – Список ip-адресов исследуемых сетей

	В контакте	Одноклассники	Фэйсбук	Google
№ AS				
ip	93.186.224.0/21 93.186.232.0/21 87.240.128.0/18	217.20.144.0/20 185.16.246.0/23 185.16.244.0/23 5.61.16.0/21	66.220.144.0/20 74.119.76.0/22 74.119.64.0/18 69.171.192.0/18 204.15.20.0/22 69.63.176.0/20	66.102.0.0/20 172.217.0.0/16 209.85.128.0/17 74.125.0.0/16 142.250.0.0/15 66.249.64.0/19 216.239.32.0/19 70.32.128.0/19 108.170.192.0/18 72.14.192.0/18 172.253.0.0/16 108.177.0.0/17 173.194.0.0/16 64.233.160.0/19 216.58.192.0/19 192.178.0.0/15
Количество активных серверов	2623	245	39	714

Количество активных серверов было определено как выборка на основании ip-адреса получателя за последний месяц. Список серверов с наибольшим объемом трафика показан в табл. 2.

Таблиця 2 – Анализ серверов исследуемых социальных сетей

В контакте		Одноклассники		Фэйсбук		Google	
ip-адрес	Трафик, Гб/мес	ip-адрес	Трафик, Гб/мес	ip-адрес	Трафик, Гб/мес	ip-адрес	Трафик, Гб/мес
87.240.182.220	135,97	217.20.145.38	8,62	173.252.100.27	11,04	173.194.21.118	28,68
87.240.182.218	127,27	217.20.157.198	7,45	69.171.247.29	4,39	173.194.21.120	28,23
87.240.182.221	121,85	217.20.145.39	7,31	66.220.152.19	4,25	173.194.21.149	27,06
87.240.182.210	114,77	217.20.153.72	6,91	173.252.110.27	2,63	173.194.21.147	25,92
87.240.182.219	106,38	217.20.157.196	5,57	69.171.235.16	0,066	173.194.21.146	24,91
87.240.182.211	101,60	217.20.153.68	5,36	69.171.246.16	0,052	173.194.21.86	24,00
Всего	13760	Всего	674	Всего	38,92	Всего	4810

На маршрутизаторе абонентского доступа одного из провайдеров Одессы с целью проведения исследования был установлен и настроен анализатор пакетов netflow и произведен анализ трафика. Результаты исследования активности использования социальных сетей показаны в табл. 3.

Таблиця 3 – Активность пользователей социальных сетей

	В контакте		Одноклассники		Фэйсбук		Google	
	Трафик, Гбайт	Запросы, пакеты	Трафик, Гбайт	Запросы, пакеты	Трафик, Гбайт	Запросы, пакеты	Трафик, Гбайт	Запросы, пакеты
1 час	106	725	3,55	831	0,25	533	31,44	789
6 часов	502,3	3844	22,2	3946	1,38	1795	180	3979
24 часа	1420	4360	74	4747	3,99	1978	532	4803
Неделя	9010	37321	441	38481	25,37	10630	3230	42719

2. Самоподобие трафика. Многочисленные современные исследования Интернет трафика свидетельствуют о том, что он обладает свойством самоподобия [4, 5, 6].

Херст параметр является мерой самоподобия или статистической инерции процесса. Оценки Херст параметра *(H)* основываются на идее измерения наклона линейного приближения на графике *log-log*. Примером такой оценки является так называемая вариограмма. График зависимости *R/S* от *N* в логарифмическом масштабе по обеим шкалам использует тот факт, что для самоподобной последовательности данных диапазон изменения масштаба или *R/S*-статистика растет согласно степенного закона с экспонентой H как функция числа включенных точек *(N)*. Таким образом, график *R/S* в зависимости от *N* на графике *log-log* имеет наклон, который является оценкой *H*.

Для вычисления степени самоподобия трафика социальных сетей будем учитывать полученные посредством протокола *netflow* данные о количестве трафика. Графическое представление данных показано на рис. 1…4, где по оси абсцисс отображен исследуемый промежуток времени (с 23 июня по 30 июня 2013 года), а по оси ординат отображено количество трафика в кбит/с, суммарное количество трафика ограничено скоростью сетевого интерфейса маршрутизатора в 1 Гбит/с.

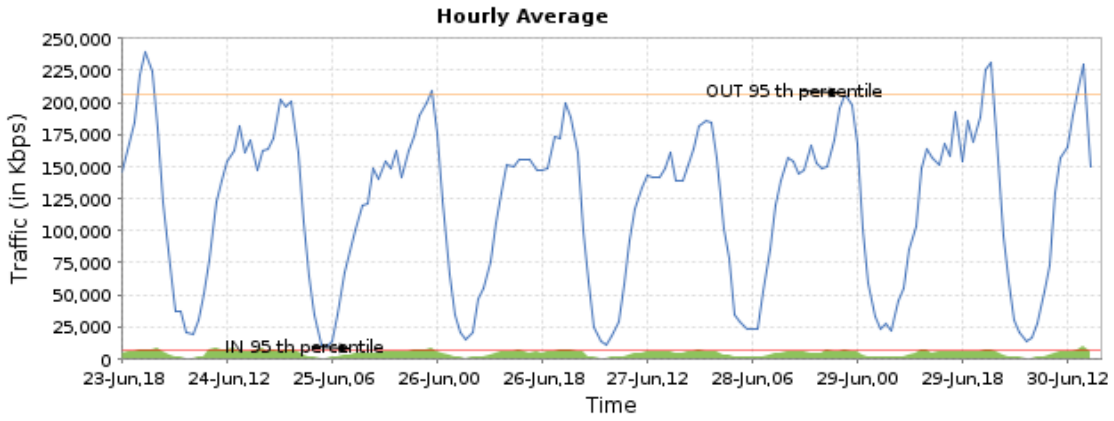

Рисунок 1 – График загрузки В контакте

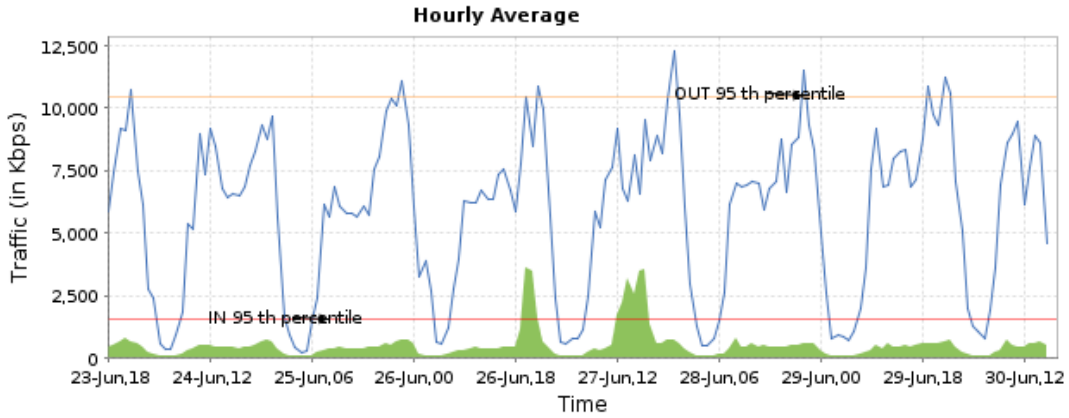

Рисунок 2 – График загрузки Одноклассники

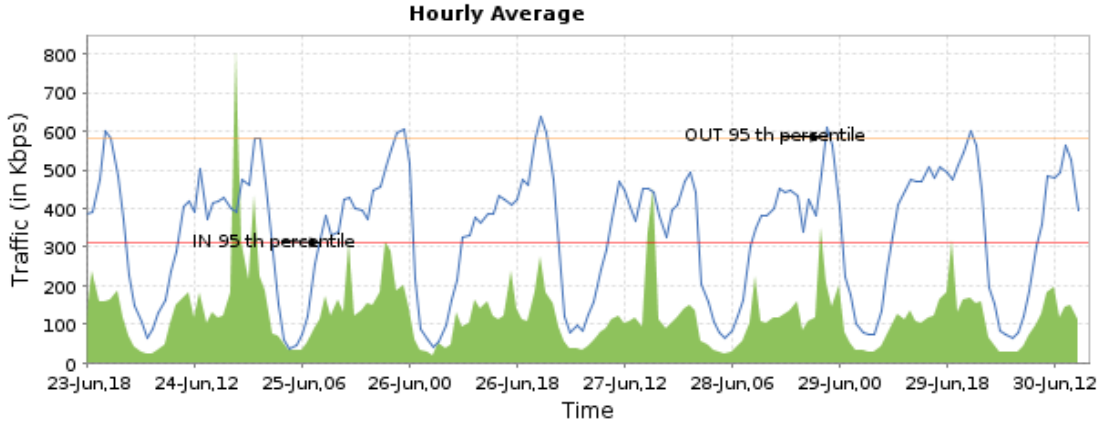

Рисунок 3 – График загрузки Фэйсбук

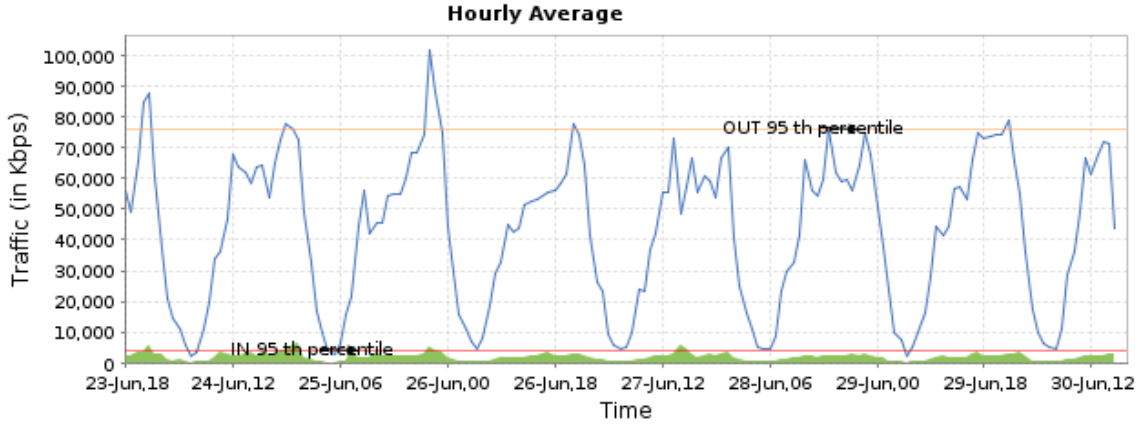

Рисунок 4 – График загрузки Google

Чтобы оценить тяжесть хвоста для имеющихся данных, воспользуемся программной реализацией методов статистического анализа MTraffic.m из программного комплекса Matlab, который позволяет загрузить подготовленные данные и получить основные статистические характеристики исследуемого трафика.

Для вычисления коэффициента Херста используем алгоритм *R/S* анализа временных рядов, реализованный в среде Matlab. В качестве входных данных используем дискретную реализацию наблюдений количества трафика $x_1, x_2, \ldots, x_N$ в соответствующие моменты времени $t_1, t_2, \ldots, t_N$, где N – объем выборки отсчетов. С учетом того, что измерения проводились каждые 5 минут, то за одни сутки $N = 288$.

Далее вычисляем накопленное отклонение для каждого из отрезков длины. Для этого обозначим через $g_j(t)$ накопленное отклонение процесса $x_i(t)$ от среднего $\overline{x}$ к моменту времени t_j :

$$g_j = \sum_{i=1}^{j} (x_i - \bar{x}). \quad (1)$$

Разность между максимальным и минимальным накопленным отклонением $g_j(t)$ определяется как размах накопленного отклонения R по формуле:

$$R = \max_{1 \le j \le N} \{g_j\} - \min_{1 \le j \le N} \{g_j\}. \quad (2)$$

Средне квадратичное отклонение определяется по формуле :

$$S = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2}. \quad (3)$$

Параметр Херста (Hurst) H определяется из соотношения:

$$\frac{R}{S} = (aN)^H. \quad (4)$$

где R – размах отклонения; S – стандартное отклонение; N – число членов временного ряда; a – константа.

Для самоподобных процессов параметр Херста оценивается из формулы логорифмированием правой и левой части отношения:

$$R/S \sim (N/2)^H, \quad (5)$$

$$H(N) = \frac{\lg(R(N)/S(N))}{\lg(N/2)}. \quad (6)$$

Используя значение показателя Херста H, выделяют три типа случайных процессов:

$0 <= H <= 0{,}5$ – случайным процесс является антиперсистентным, или эргодическим рядом, который не обладает самоподобием;

$H = 0{,}5$ – полностью случайный ряд, аналогичный случайным смещениям частицы при классическом броуновском движении;

$H > 0{,}5$ – персистентный (самоподдерживающийся) процесс, который обладает длительной памятью и является самоподобным [9].

Коэффициент пачечности для заданного процесса соответствует отношению пиковой интенсивности процесса поступления заявок на обслуживание к его среднему значению. Самоподобие можно расценивать как фундаментальное статистическое свойство сетевого трафика, которое необходимо учитывать на практике.

Полученные результаты показателей Херста для исследуемых социальных сетей сведем в табл. 4

Таблиця 4 – Показатели Херста исследуемых сетей

	В контакте	Одноклассники	Фэйсбук	Google
Понедельник	0,75	0,73	0,71	0,75
Вторник	0,78	0,76	0,72	0,79
Среда	0,73	0,75	0,72	0,74
Четверг	0,71	0,80	0,69	0,76
Пятница	0,75	0,77	0,70	0,77
Суббота	0,78	0,78	0,72	0,79
Воскресенье	0,81	0,77	0,72	0,8

На конец стоит отметить, что исследования самоподобных свойств трафика позволяют с достаточной степенью достоверности прогнозировать появление на сегменте сети временных периодов с перегрузкой по производительности оборудования и линий связи, что, в свою очередь, делает возможным построение системы с динамическим управлением пропускной способностью для отдельных видов трафика. Кроме того, подобное прогнозирование используется при разработке алгоритмов, направленных на повышение качества обслуживания.

В заключение можно сказать, что поставленные цели и задачи исследования достигнуты. Проведенные исследования количественных и качественных характеристик наиболее популярных в Украине социальных сетей, позволяют сделать следующие выводы:

- доля трафика социальных сетей является весомой, постоянно увеличивается и на июнь 2013 года, для исследуемой сети провайдера составляет до 25 % от общего объема;
- социальная сеть В контакте является лидером по количеству трафика и активных серверов для пользователей Украины и опережает таких гигантов, как все сервисы компании гугл;
- трафик социальных сетей – самоподобный процесс;
- коэффициент самоподобия трафика меняется от 0,69 до 0,81;
- измерения показателя Херста показали, что для всех видов трафика социальных систем $H > 0{,}5$, трафик относится к классу персистентных процессов.

Наиболее перспективным направлением дальнейших исследований представляет интерес анализа различных видов трафика методами нелинейной динамики. Для повышения эффективности распределения самоподобного трафика компьютерных сетей необходимо создание математических моделей, которые наиболее полно отражают фрактальные свойства процессов.

Литература

1. *Губанов Д. А.* Социальные сети: модели информационного влияния, управления и противоборства / Губанов Д. А., Новиков Д. А., Чхартишвили А. Г. – М.: Издательство физико-математической литературы, 2010. – 228 с.
2. *Градосельская Г. В.* Анализ социальных сетей: дис. ... кандидата социологических наук: 22.00.01 / Градосельская Галина Витальевна. – М., 2001. – 223 с.
3. *Зверева У.* Исследование аудитории российских социальных сетей [Электронный документ]: Аналитический портал веб разработок / У. Зверева, М. Здановская – Режим доступа: http://research.cmsmagazine.ru/audience-research-russian-social-networks/.
4. *Петров В.В.* Структура телетрафика и алгоритм обеспечения качества обслуживания при влиянии эффекта самоподобия: дис. … кандидата тех. наук: 05.12.13 / Петров Виталий Валерьевич – М., 2004. – 199 с.
5. *Зюльков И.А.* Характеристики самоподобия случайных процессов и трафика радиосистем при наличии повторных сигналов: дис. … канд. физ.-мат. наук.:01.04.03 / Зюльков Илья Александрович. – Воронеж, 2004. – 161 с.
6. *Цыбаков Б.С.* Модель телетрафика на основе самоподобного случайного процесса / Цыбаков Б. С. // Радиотехника. – 1999. – № 5. – С. 24-31.
7. *Леинванд А.* Конфигурирование маршрутизаторов Cisco. – [2-е изд.] / А. Леинванд, Б. Пински – М.: Издательский дом "Вильямс", 2001. – 368 с.
8. *Уильям Р. Паркхерст* Справочник по командам и настройке протокола BGP-4 маршрутизаторов Cisco / Уильям Р. Паркхерст. – М.: Вильямс, 2002. – С. 384.
9. *Заборовский В.С.* Сети ЭВМ и телекоммуникации исследования процессов в компьютерных сетях: телематический подход / Заборовский В.С., Мулюха В.А., Подгурский Ю.Е. – СПб.: Изд-во СПбГПУ, 2009. – 159 с.