Big data to Big Insight?

Turning the noise of social media into actionable intelligence

The pace of change in the world appears to be increasing, and technology is driving that – or at least exposing it. More people are using internet-connected devices to stay informed, communicate and run their lives. For "millennials", the internet, mobile devices and social media are not regarded as transformational technologies, but a normal part of life. In this context, new forms of media such as social media, user generated communities and citizen journalism are an essential source of rapidly-disseminated information about what is actually happening in the world. In many cases it is driving change, but at the very least it is now an important information source for any organisation that needs to monitor the world for emergent trends and risks.

Many organisations have developed their first generation of social media monitoring approaches. These strategies range from simple searching for keywords on websites, to automated systems that monitor, collate and alert users to relevant social media data. What is lacking in many of these tools is the ability to provide insightful answers, rather than presenting the user with yet more data. It is not enough to simply have a monitoring 'dashboard', which may look impressive but in fact provides an illusion of intelligence. We speak to many teams who report that they are already covering social media, but on further reflection reveal that their existing tools and service providers have limited value exactly because they fail to deliver intelligence as opposed to information. Without understanding the strengths and limitations of the data and the models it is impossible to weight such information in critical decisions. Social media, like many other data sources, is messy, incomplete, biased and confusing. The next generation of systems will need to perform better at collating the right information, structuring and visualising key insights, and adopting an 'augmented intelligence' approach to combine data and algorithms with human expertise.

For the past year analysts and experts from IHS Country Risk , IHS Aerospace, Defence & Security, and IHS Energy have been collaborating with the data and network scientists from JANYS Analytics to develop methods and tools to gain insight from social media. Two key themes have emerged from this work. Firstly, data and analytics alone generate insights with minimal value. Meaningful answers only emerge when subject-matter experts are working closely with technical specialists. Secondly, statistical models aimed at predicting complex social and political events can only be one input amongst many for an analyst. Rather than trying to build sophisticated predictive models, analysts can start to gain insights from much simpler visual analysis tools with a lower empirical bar – layering up time series, geospatial and network analytics visualisations to identify patterns and trends. This approach can have its own risks, in particular of seeing patterns where they don't exist, or inferring meaning from invalid data. But used appropriately it can trigger valuable insights and give early warning of new trends and issues.

Drawing from this experience, this paper outlines some of the key challenges in designing a successful Social Media Intelligence (SOCMINT) strategy for corporate, risk, security and strategy teams, highlights some of the main areas where social media can provide useful information, and raises questions that should be asked by decision makers designing and resourcing a SOCMINT system. The starting point, as always with such matters, is to have a clear understanding of what questions you are trying to answer, and a clear operational concept of how social media should fit into existing analytical and decision processes. Most importantly, decision makers should not be reactive and base strategies on a perceived need to do something with social media, but rather have in mind the opportunities that this domain can provide for their organisation.

Know your platforms: and which are relevant.

The variety of social media platforms is bewildering and growing. Without the knowledge of country subject matter experts, it is difficult to start to identify which types of platform are relevant for your specific questions. Platforms typically fall into one of four categories:

1. Social Networking – sites which enable personal connections based on friendship, work or shared interests, such as Facebook or LinkedIn. Facebook is the most popular social network in just about every country worldwide (and in many cases is the most frequently visited site – topping even Google). The major exceptions are China (where Qzone has 600million monthly users) and the former Soviet states where Vkontakte in particular is popular.

2. Blogs & Microblogs – services that provide a platform for users to share thoughts and ideas – blogging has been a long established internet use and while it has faded somewhat in the US and Western Europe, sites such as Blogger are hugely popular in the Middle East and Livejournal is in Russia. Increasingly Twitter, with its 140 character limit, has replaced long-form blogging. Its Chinese equivalent is Sina Weibo.

3. Content Communities – websites focused around the creation and sharing of media such as photographs and videos. As with all social media this ranges from the irreverent to the overtly political and inflammatory.

4. Instant Messaging – services that allow for rapid, free, private chats and sharing of media between individuals over the internet, often through smartphones.

However, increasingly the uses and functionality of these platforms are blurring - for example Twitter is becoming a social network as much as a microblogging platform, and Facebook is increasingly an instant messaging platform. Pictures and videos hosted on a content site like Youtube can rapidly spread around networks of individuals over instant messaging services. The way that people use platforms also evolves and categorising them becomes less meaningful – for example Sina Weibo, ostensibly a Chinese equivalent to Twitter, where 140 characters in Chinese is more like 80 words in English, which means it is used in very different ways. Such regional differentiation in use of platforms in large part determines where a social media intelligence system should be focused.

Social Networking	Blogs & Microblogs	Content Communities	Instant Messaging
Facebook	Twitter	You Tube	WhatsApp
Qzone	Sina Weibo	Pinterest	Blackberry Messenger (BBM)
Linkedin	Blogger	Instagram	Facebook Messenger
Kohtakte	LiveJournal		Tencent QQ

Know your platforms: what apps matter where?

The use of social media platforms varies markedly from country to country. There is wide differentiation in internet penetration both across regions and within countries. Whereas nearly 90% of Bahrain's population has internet access, less than 10% of Iraq's does. Users of the internet are generally skewed towards younger, wealthier and urban populations, so any analysis of social media has to be cognisant of the people the data is representing – and those that are excluded that are therefore not going to be sampled in social media data.

The platforms that are used and popular also varies a lot between countries. For example, across the Middle East region it is Facebook that is the most popular social media site by total traffic, with Youtube following. It is only in Turkey, Saudi Arabia and Lebanon where Twitter is in the top three. This has implications for how easily analysts can monitor relevant risks, as by its nature Facebook is a more 'closed' network than Twitter and more difficult to analyse at scale.

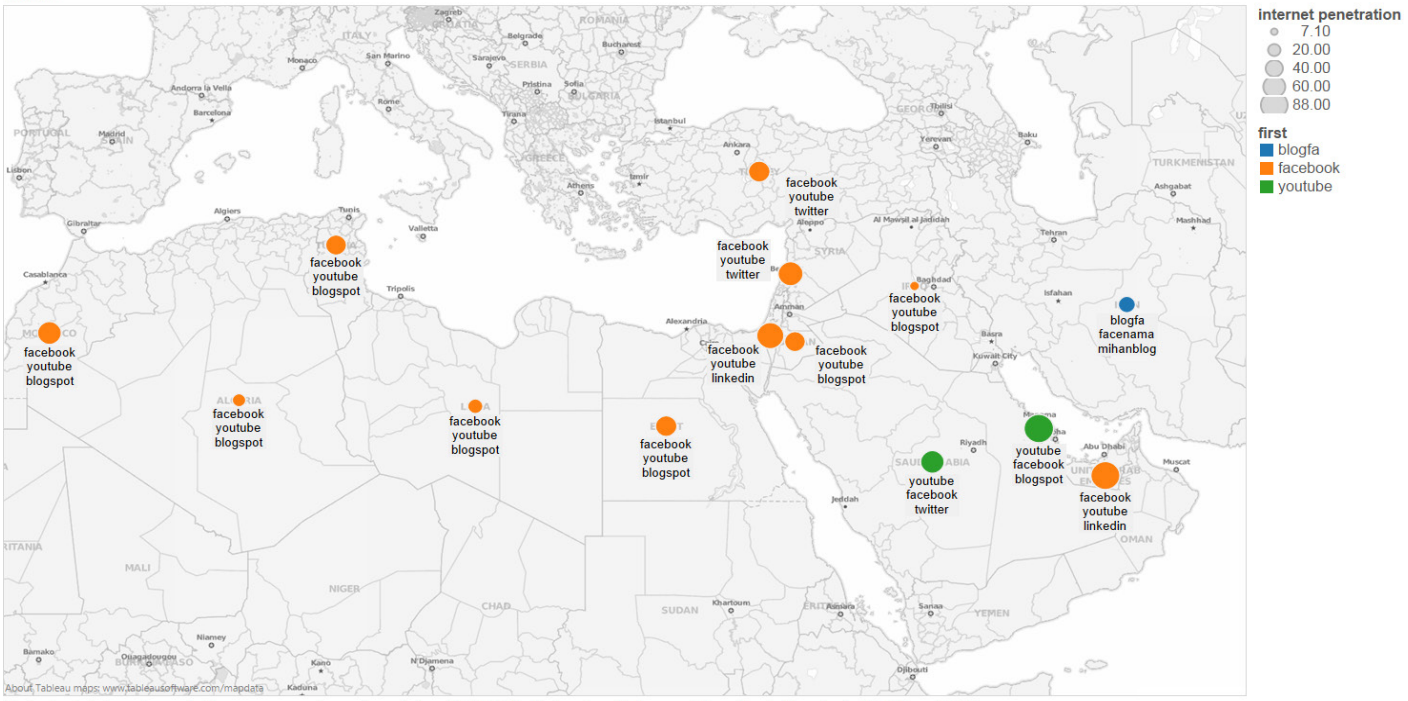
Any SOCMINT strategy therefore needs to be cognisant of how social media is used in the countries of interest and adopt methods that allow the right platforms to be assessed.

Know your platforms: who uses them, and how?

If it is important to understand which sites are used where, it is as important to understand what sites are being used for in these locations. Internet traffic data alone does not tell you how different groups use different social networks for different purposes. You cannot figure that out by looking at hit data, only by deeply engaging with the content and messaging on these platforms in local languages. If we look at the MENA region again we can see that use is very differentiated:

Twitter is the primary platform for political debate in Egypt and Tunisia. It's also used in Saudi Arabia to openly discuss politically sensitive issues with relative anonymity – an important data source on political issues given that the country has reportedly the highest number of Twitter users relative to internet penetration anywhere in the world.

Internet penetration rates and top social media platforms in MENA region. Source: ITU, Alexa



Mark **size is internet penetration** as a percentage of population (ITU, 2012). Mark **colour is top social media platform** by traffic (Alexa, 2013).

Facebook is the main platform in Algeria used by workers' unions and political movements to organise their supporters and announce upcoming strikes and protests. In Bahrain, the radical February 14 Youth Coalition uses Facebook to document violent actions and operations by the security services. A number of North African jihadist groups also use Facebook to publish official statements, photos and videos of their activities. In Syria, there are well over a thousand rebel groups, and in nearly all cases they have their own facebook page with a founding video and details of claimed operations.

YouTube is used extensively by Shia youth groups in Bahrain to document anti-government protests, in conjunction with Facebook. These are usually not reported anywhere else. Furthermore, the videos can often be attributed to an exact location, which can be geocoded and used to create extremely detailed maps of protest hotspots. It is also used frequently by insurgent groups in Syria to showcase new weapons systems and document military advances. AQAP in Yemen has published YouTube interviews with local commanders discussing their ideology and tactical details of attacks against the military.

This variety of use and media types reinforces the need to be clear about what questions you are trying to answer. If it is to gauge the mood of the Saudi population to political issues, Twitter is probably the best data source. If the objective is to understand what weapons systems different rebel factions in Syria are using and how this is evolving, Youtube will be of more use. But using different platforms requires very different tools and approaches.

Know your data: are you getting the full story?

If you are using an external platform to search and plot social media content, you first of all need to understand what data your provider is accessing. The full Twitter 'firehose' provides streaming of all public tweets – in the region of 450 million a day. The 'decahose' provides a 10% random sample of this. The public streaming API that Twitter provides for free has a host of limitations, and is always capped at 1% of all tweets, so any monitoring products based on the latter will have very limited validity.

Most social media platforms are much more limited than Twitter and do not provide such easy access to detailed data that can be used for modelling and analytics. Many monitoring approaches for Facebook, for example, are based on public 'pages' or 'walls'. While you may be able to see how many 'likes' a page (such as a protest campaign) has, this does not directly translate into whether people are likely to take real-world direct action for a cause (activism vs slacktivism). It does not give you data on the social graph of individuals engaged on an issue, their relationships and networks of influence, or how information is passing between them. Many services do not provide any data at all to resellers for use in monitoring applications – in particular private chat or instant messaging applications. This highlights the importance of having a clear and ethical monitoring strategy based on data that is legitimately acquired from open sources. Furthermore, recognising the scale and importance of gaps in information is necessary in order to be able to appropriately weight the relevance and value of what you can see.

This also reinforces the importance of traditional human intelligence – having people on the ground, in the population, who pick up on emergent issues and discourses and can provide awareness and context. They can tell you when something has changed, when all their friends who are normally apolitical start passing on calls to protest, when cafes and dorms are buzzing with instant messages of rumours of police brutality. HUMINT in this context can be an essential bellwether for what is happening on the web rather than vice versa.

Know what is accurate: verification amid the volume

When assessing the most viral tweets related to the April 2013 Boston marathon bombing, IBM Research Labs found that nearly a third of the tweets were rumors or fake content. Half of the tweets were merely generic comment, and only a fifth of tweets had true information about the attacks.

Moreover the researchers found that even users with verified accounts and long standing good 'social reputation' and many of their followers were quick to spread false information without verifying it. In a rapidly developing emergency, reaching to social media for situation awareness represents significant challenges for information verification and validation, but that doesn't mean it cannot be useful. If relying on social media for specific information about what is happening in a situation it is important to track back to its source. Once that is done, source assessment techniques can be applied, assessing characteristics such as how long an account has existed for, and its pattern of interaction with other users.

By assessing a source's value and reliability using a range of standardised criteria – not dissimilar to other source evaluation - it is possible to cut down the 'noise' on social media platforms and focus monitoring on users that you have high confidence in. Such an approach allows analysts to deploy collection resources most efficiently on the sources that matter and weight new information appropriately.

This is particularly important as social media becomes a 'contested commons' where actors are using platforms for disinformation. Governments are carrying out disinformation campaigns on social media, and activist groups are well aware that their public discourse is being monitored.[1]

By filtering down to a qualified list of sources for each country, a regional intelligence team can use monitoring tools to track the tweets of those users in a structured manner, complementing other collection methods. This again highlights how a successful approach to social media is predicated on application of traditional intelligence tradecraft.

Why Tradecraft and Analytics both matter

Intelligence Tradecraft	Data Analytics & Network Science
Asking the right questions – 'setting the dials'	Collecting, processing and filtering massive volumes of data
Identifying and monitoring all the right platforms & sources	(Semi) automated monitoring of key sources where data feeds are available
Pattern recognition (joining the dots, identifying significance of data)	Visualisation, mapping, and alerting of trends in data
Assessing relevance of actors on real-world risk outcomes	Using algorithms to suggest key influencers and groups
Combining all-source inputs with experience and context to produce specific, tailored risk forecasts.	Delivering an intelligence product in interactive, intuitive formats.

While this approach does not require extensive proprietary technology, it does require time and resources to build up the knowledge base of which sources are valuable and reliable, and keep that list up to date. For corporate security leaders, using social media smartly to ensure that it adds value rather than detracts by taking up limited and valuable analyst time is crucial. This will always be easier for larger teams with greater collection resources, but raises questions about the role of the analyst in an organisation and whether their main focus should be collecting intelligence or interpreting it for the business. As data access increases exponentially, maintaining an effective focus on the latter will become ever more challenging and require more systematic approaches to filtering and allocating resources.

1 Lots of examples of this (including S Korean intelligence being accused of spreading 1.2 million tweets to influence elections).

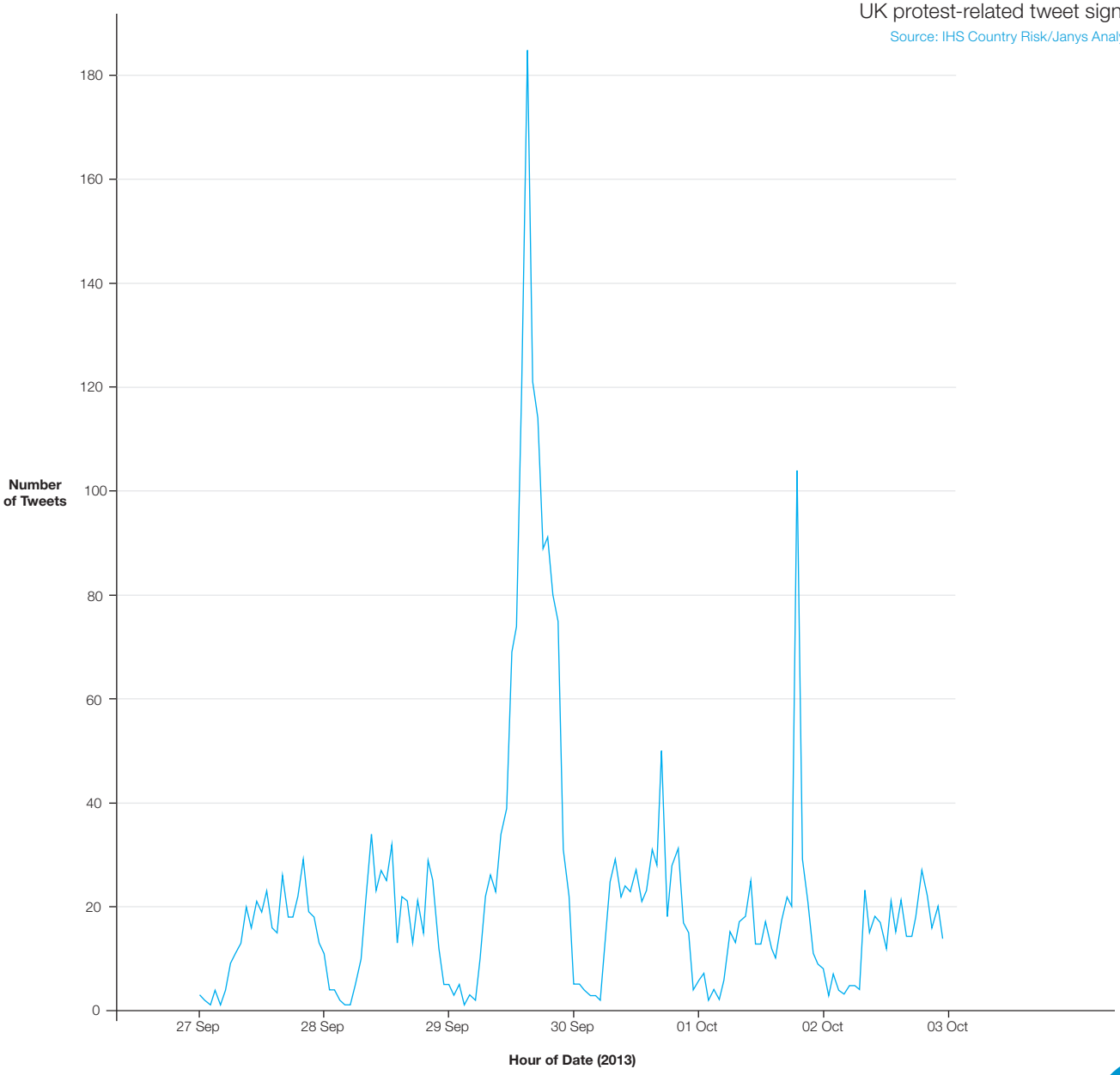
Know what can be predicted: Social media's utility in detecting events

With the right system, and the right data, it is possible to use SOCMINT to get early warning on emergent events ahead of what would be practically achievable with a traditional media monitoring system. There are now a number of cases where research has shown that the signals in social media reduce the latency between events on the ground occurring, and them being visible to analysts. This shouldn't be conflated with prediction, however.

The challenge with all such intelligence systems is calibrating them to balance signal and noise. Which is more important – to ensure that no signals are missed (and therefore increase the rate of false positives), or to avoid overloading the user with noise (and subsequently tighten the system's filter, increasing risk of false negatives – or in other words, missing something important). There is no right answer here, and decisions need to be made about what is acceptable, and what resources are going to be allocated to running the system. If your team is a handful of people, is it realistic to expect them to be able to gather all pertinent information to global risks from their desktops?

UK protest-related tweet signals

Source: IHS Country Risk/Janys Analytics



Know what algorithms can do: Analytics are an input, not the answer

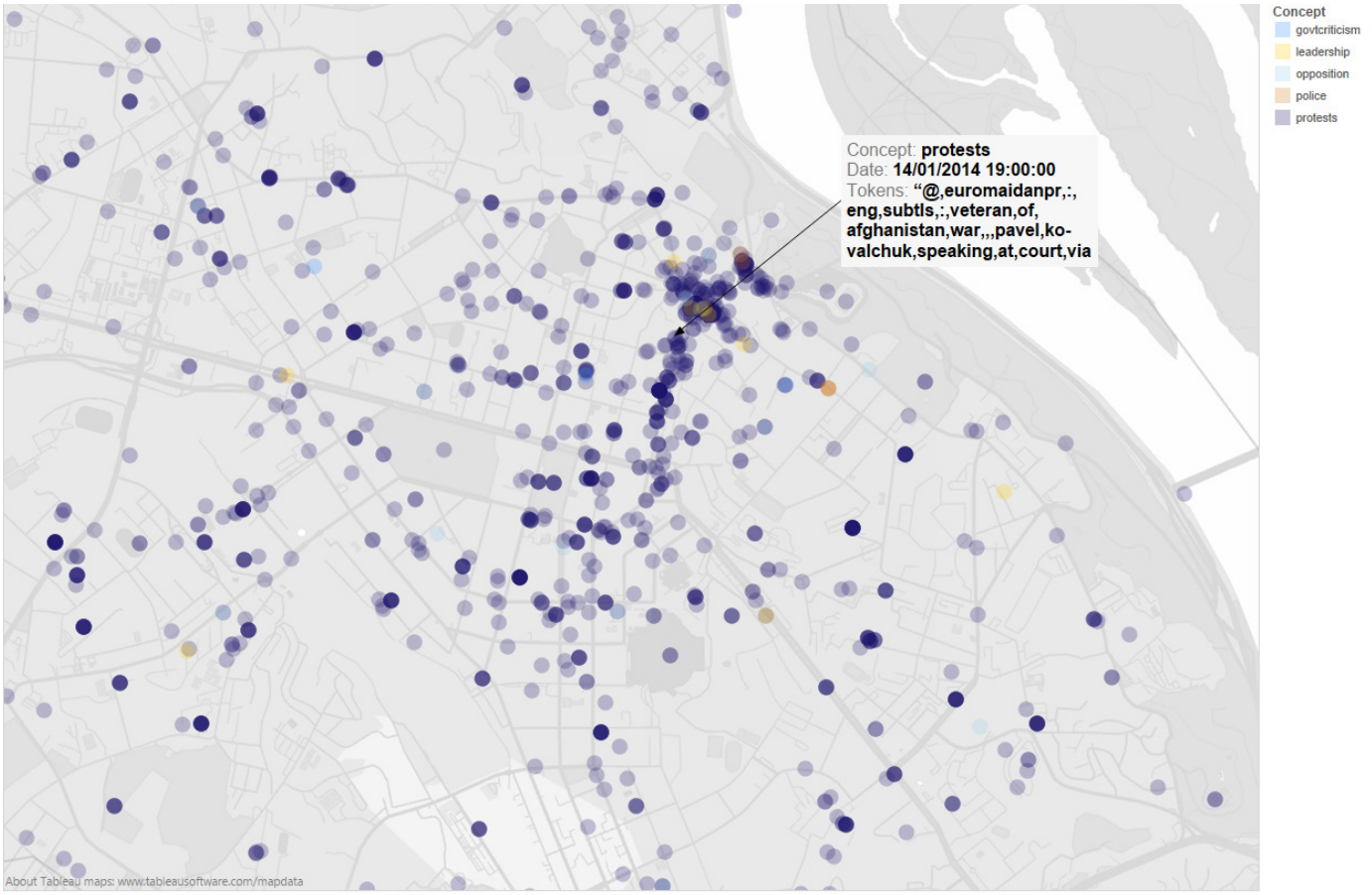
Most social media tools have some form of sentiment measurement built in, intended to measure the mood of a user in their posts. These methods often stem from marketing and brand management applications – allowing a customer service team to identify users who are particularly angry about their company, for example, and then reaching out to communicate with them directly. Applying such methods to complex socio-political issues is much more challenging. Measuring the mood of a population and for the purposes of predicting social unrest is probably unrealistic. Validation of sentiment measurement methods has shown that the wider the sample (in time or population), the less meaningful the signal. In other words, any trends get averaged out, or hidden amongst factors such as weather or sports results. When assessing the reach and influence of the Somali militant group Al Shabaab on twitter, our analysts found that automated sentiment methods did not tell us anything useful, but when we used text mining to uncover prevalent themes we were able to quickly get a qualitative assessment of the opinions towards the group (highly negative, and focused on calls for the group's banning on Twitter). Where sentiment models seem to be useful is answering a very narrow, targeted question, such as identification of users who appear, based on their posts, to be the most negative on a particular issue, such as fracking. When combined with other analytics tools (community detection, influence measurement, geolocation), it can help narrow down the universe of billions of users to a manageable number of key people worth listening to. It does not help you predict, but it does help you make resourcing choices.

Know where signals matter: Geolocation of social media data is limited

Currently only around 1% of all tweets have latitude and longitude data, sent when people opt-in to sharing their location on smartphones. As mobile usage increases this is expected to increase. This information can be used to assess where people are tweeting about certain issues, such as protests. By aggregating the data over longer time periods it is possible to see hotspots of activity, and detect changes from normal patterns. A large number of people suddenly tweeting in a certain district of a city may be a cue for an analyst to investigate further what is driving that behaviour.

The majority of social media data is not so readily spatialised. On Twitter for example, unless the user is transmitting their coordinates, the only way of inferring their location is through their language or their self-reported location in their profile page. This can be parsed and used to filter tweet data by country, city or state, but is not an exact science, particularly as there is no way of knowing whether their profile is accurate, and it is not mandatory to fill in. In future models to estimate location based on the topics and networks of individuals are likely to improve the proportion of data with good quality location information, but currently it is a major gap.

Geolocated tweets in Kiev, Ukraine, January 2014



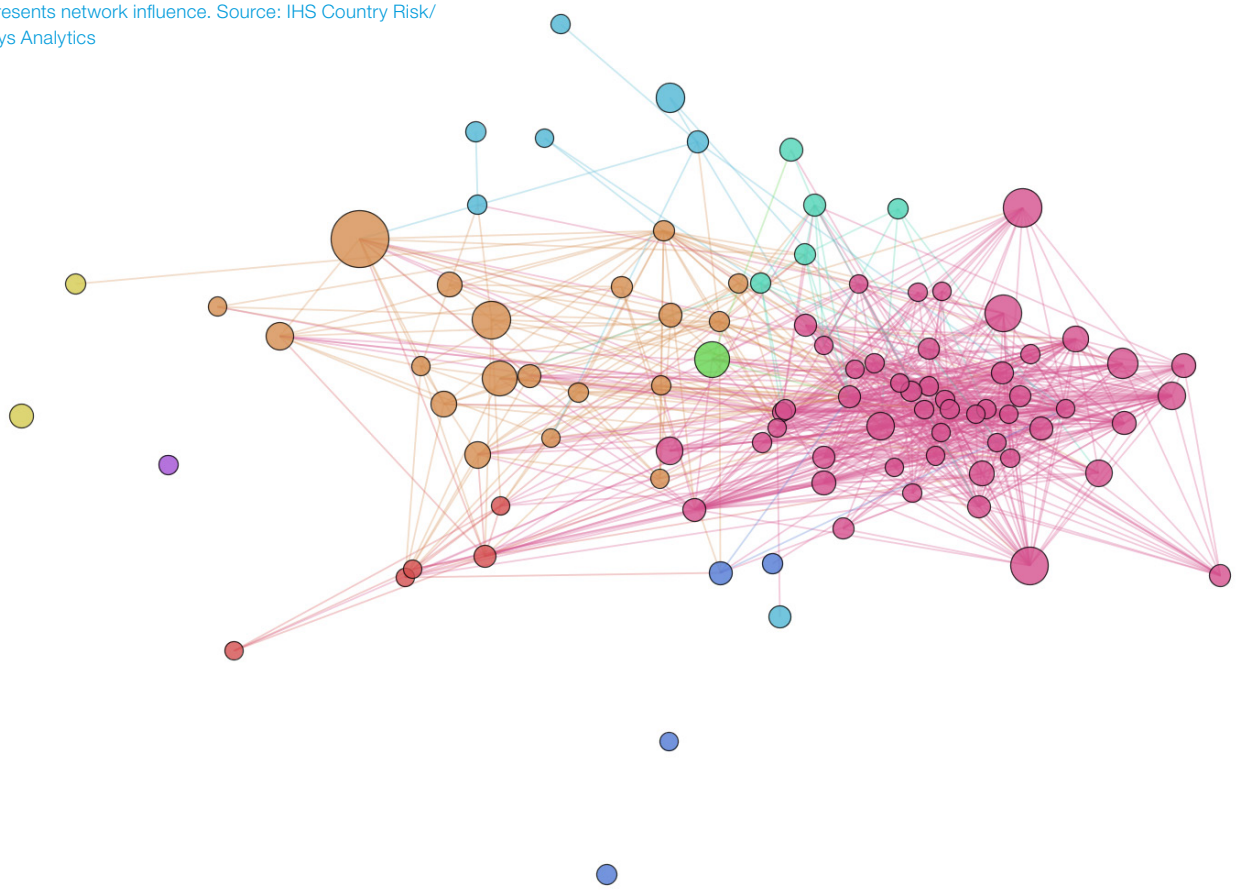
Map based on Long and Lat. Color shows details about Concept. The data is filtered on Month, Day, Year of Date, Day of Date and Country. The Month, Day, Year of Date filter excludes February 22, 2014. The Day of Date filter includes dates on or before 22 February 2014. The Country filter keeps Null and ukraine.

Know how to make sense of complexity: Mapping of leaderless movements using network structure

Social media can provide data to rapidly make sense of the structure of complex events and movements in a way in which traditional sources of information (new reporting or well-placed individuals) cannot. During the anti-government protests in Turkey in June 2013 we gathered a dataset of over 3.5 million tweets related to the protests. A key dimension to these protests was their leaderless nature. They were the biggest anti-government protests since the AKP party came to power, but nobody was directing them. While opposition parties were involved, they were not controlling the protests. As a security analyst with executives or operations on the ground in Istanbul, how do you begin to understand where a leaderless movement is heading? Will protests spiral out of control, turn violent, spread elsewhere?

We ran 'community detection' algorithms over the social graph of people tweeting about the protests. These look for clusters of users who are more closely connected with each other than with users in the rest of the network (because, for example, they retweet each other more often). For a corporate security analyst, trying to make sense of a complex and dynamic situation, this is powerful information. It gives you a sense of how the communities map onto real world groups, such as student activists, pro-government supporters, or Turkish diaspora. It thus helps you understand who the actors are, and helps you ensure your analysis is covering all the bases it needs to.

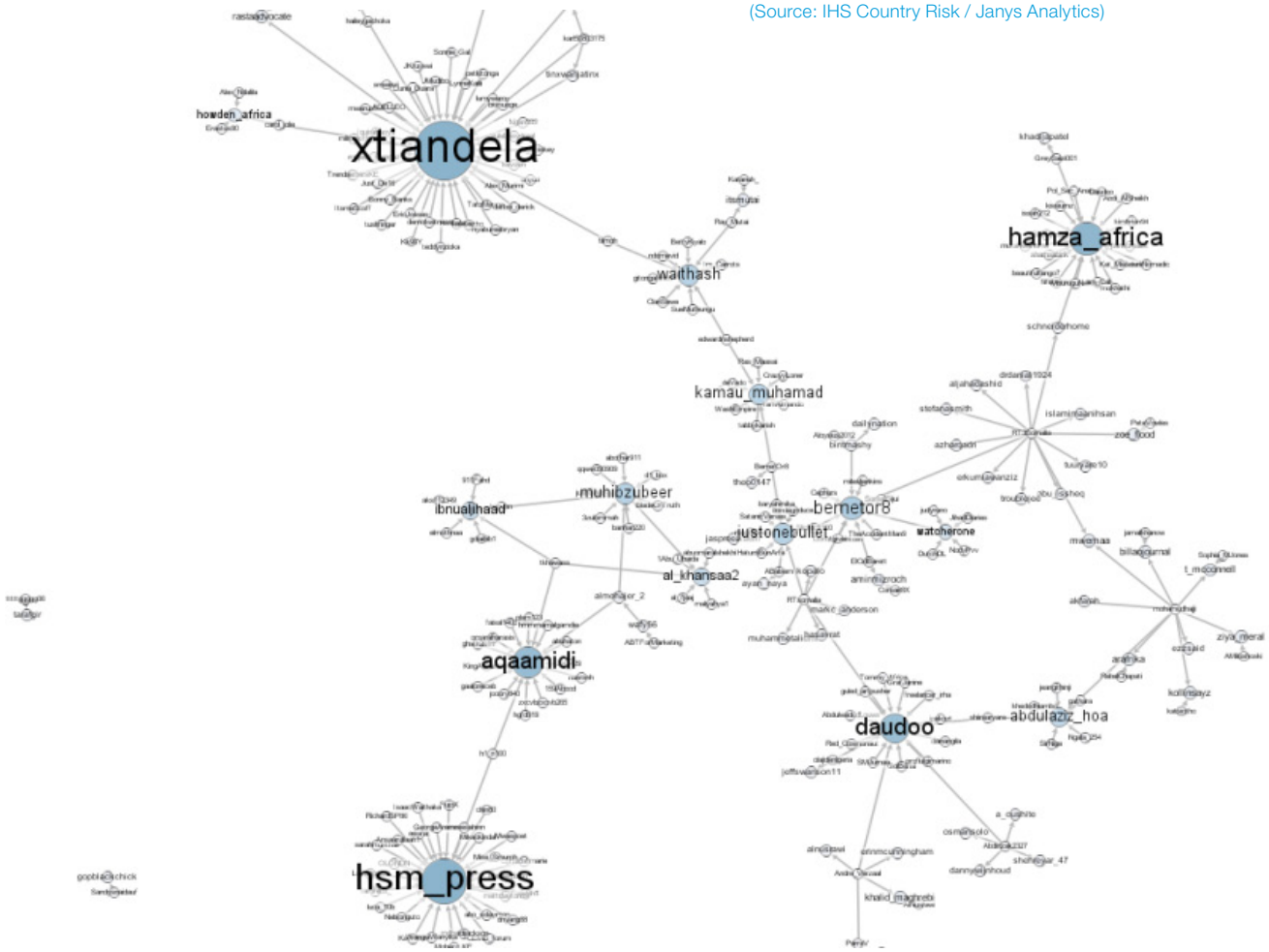
Communities of users tweeting on Argentina oil field developments. Node colour signifies community, size represents network influence. Source: IHS Country Risk/ Janys Analytics



Know who is important: Network analytics can help identify key influencers

The structure of a network can then inform an understanding of who is most influential within it. When assessing the reach and influence in social media of the Somali militant group al-Shabab, we constructed a network of Twitter users who retweeted messages from al-Shabab in the period before and after the Westgate siege in Nairobi in September 2013. We consider retweeting to be the most influential type of Twitter relationship because it is a direct repetition or endorsement of another user's message. Following is mostly a superficial relationship: few (if any) of the millions of users who follow Barak Obama have any direct interactions with him over twitter (direct interaction meaning a mention, or Obama retweets you, or replies directly to your message.) In the Kenya case,we found that the most influential people were in fact journalists and bloggers, suggesting that al-Shabab's messages were perhaps reaching the right people from their perspective in order to spread awareness.

What this technique enables an analyst to do is filter down to who really matters, within the huge numbers of people who may be commenting on a particular topic. This information can then be used to focus monitoring efforts (ie tracking what key influencers are saying), and in other cases working with stakeholder engagement and external affairs teams to reach out to these influencers. A pro-active strategy of managing an issue before it causes a problem, for example by meeting with a key local activist leader to discuss their concerns before they start to mobilise action against a new project, is better than reacting once the threat has emerged.



Al-Shabab retweet graph, node size indicating network influence (Source: IHS Country Risk / Janys Analytics)

Know how to join the dots: Pattern recognition and the intelligence cycle

Putting together all these capabilities, how can an organisation set up a monitoring system, then join the dots to identify significant and risk relevant developments amid the noise? One technique, type of output or model cannot tell you everything you need to know. This is why it is necessary for an analyst to have access to systems that can reveal various aspects of a situation, in an intuitive 'visual analysis' environment. Sometimes the geospatial map will be most useful as it will tell you there is an unusual concentration of activity in a certain area of a city, but then qualitative text analytics of the concepts that are being tweeted about is needed for the analyst to decide if the activity is risk relevant or not (are people tweeting because a major music act is in town, or because there have been reports of police brutality)? Then you use network analytics to determine the urgency of the signal (is it key influencers who are calling for protests, or people who are active online but do not have the ability to actually get protestors on the streets)?

Crucially, analysts need to understand what is actually a valid inference to make from the data. For example, a significant spike in activity may be evident in a time series, but if the Twitter user base in a country is very low and biased towards a particular demographic, that trend may not be meaningful. Analysts also need to be able to rapidly experiment, to test hypotheses against new combinations of data – many of these hypotheses will be wrong, but with a dynamic system, such learning will have low opportunity costs. Any SOCMINT system will produce many false positives and many false negatives, so it is critically important that it is part of a wider intelligence collection and analysis process – to make sense of what is happening, and to fill in the gaps that social media will not cover.

We believe that any successful SOCMINT system has to be developed in close partnership between subject matter expert analysts, data scientists, and software engineers. Iteration and improvement cycles are critical, as are having realistic expectations at the start with regards to what a system can provide given the available resources. We look forward to learning together with our customers as we develop unique capabilities in partnership.

Strengths and limitations of SOCMINT

What SOCMINT can do	Limitations of SOCMINT
Enhance situational awareness of fast-moving complex developments (e.g. protests, coups)	Only a portion of the drivers of global risks are observable through social media (users are relatively young, educated and connected)
Information on events poorly reported in other sources (e.g. due to state control of media)	Most data related to networks and communications between individuals is not publically available – or cannot be easily analysed at scale.
Measure the scale and sentiment of online discussion of a topic of interest (such as a direct action campaign), and gather data on 'atmospherics' in a population.	Algorithms can only go so far in providing answers to questions about complex risk environments. Human analysts and decision makers are central.
Provide timely (or early) warning of signals of emergent risks (one set of indicators amongst many) – such as growth of a new protest movement	Signal to noise ratio is very low, and any system will struggle to catch all indicators without overloading end-users with noise.
Identify key influencers and groups of users to inform monitoring and outreach (the people who are most likely to influence risks)	

SOCMINT strategy checklist

Invest in training – learn what works and what doesn't and how to interpret data, before you invest in systems.

Define an operational concept – what you want to achieve, and what is possible now and in the future. What is the role of internal analysts, and what should be outsourced?

Start small, develop a portfolio of tools and methods that work across multiple platforms and can address the concerns of different teams within your organisation.

Iterate and build. Keep an agile mindset and keep learning, as the platforms and tools change.

For further information about social media intelligence in IHS Economics and Country Risk, please contact:

Dr Nathalie Wlodarczyk
Managing Director, Global Risk Consulting
IHS Economics and Country Risk
+44 20 8276 4736
nathalie.wlodarczyk@ihs.com

David Hunt
Senior Manager, Risk Indicators & Analytics
IHS Economics and Country Risk
+44 (0) 20 8276 4717
david.hunt@ihs.com