USC EE 542

Lectures 23 and 24, Nov.13, 15, 2017
Professor Kai Hwang, USC

- **Machine Learning Model Fitting (Sec.6.4)**
- **Social Network Clouds (Sec.5.3)**
- **Social Graph Analytics (Sec.9.4)**

Basic Concepts of Machine Learning Performance

- **Machine Learning Performance Metrics**
  - **Quality of Training Datasets and Training Time**
  - **Accuracy of Forecast, Prediction, and Classification results**
  - **Implementation, Update, and Maintenance costs**
- **Sampling Scores vs. Cross Validation Scores**
- **Model Fitting Modes:**
  - **Perfect Fitting (Sampling score matches with Testing scores**
  - **Overfitting (Model over biased by sampling data)**
  - **Underfitting (Poor choice of sampling data)**

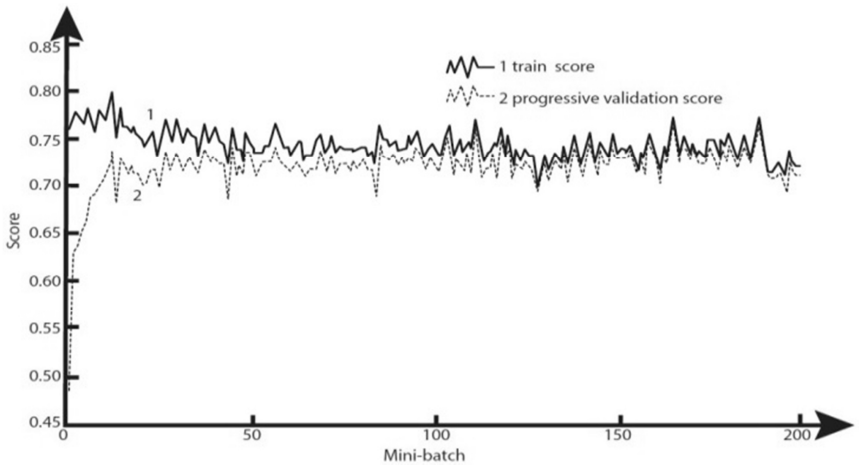


Figure 6.19: The training score and cross validation score match nicely in a well-fitting machine learning model created.

Selecting Data Sets or Training Samples

- **Common Datasets: Dividing the original data set into two parts: with equal characteristics and distributions in the training set and testing set. This subdivision should result in a model performance to avoid either the over-fitting or the under-fitting problems.**
- **Cross Validation: Dividing the original data set into k parts and selecting a part in turn as a test set while the remaining as the training set. This demands k validation testing runs. This model performance shows the average accuracy of the model over many subdivided test sets.**
- **Bootstrap Cycle: Randomly sampling with replacement of some data elements repeatedly in different training samples. Let the sampled data be the training set, while the remaining as the test sets. Repeat these sampling cycle k times. The may end up with a weighted mean performance of all test sets.**

Selecting Algorithms for General Datasets

- Given a dataset from a known application domain. The following procedure shows how to select the the proper machine learning algorithm, based on the dataset characteristics and performance requirement. Consider six ML categories: Decision Trees, Regression, Clustering, Bayesian, SVM, and ANN's.

- In general, the following options can be considered in solving the under-fitting problems. These methods appeal in particular, to improve the classification problem Since the model performance is so sensitive to the data sets applied. We consider three options to select the data sets.

Over-Fitting in ML Models

- This is the case that the training score is very high, but the cross-validation score is very low for testing datasets applied.
- As shown in Fig. 7.28, the two scores are separated far apart from each other. This status implies that the model fits the training set very closely.
- Overfitting model has ignored the noise margins in the validation data set. In other words, the training set is heavily biased on a particular training data set.
- This sample data set stays far away from common data distribution or characteristics in general applications. In this case, the overfitting model simply cannot model the testing data, accurately.

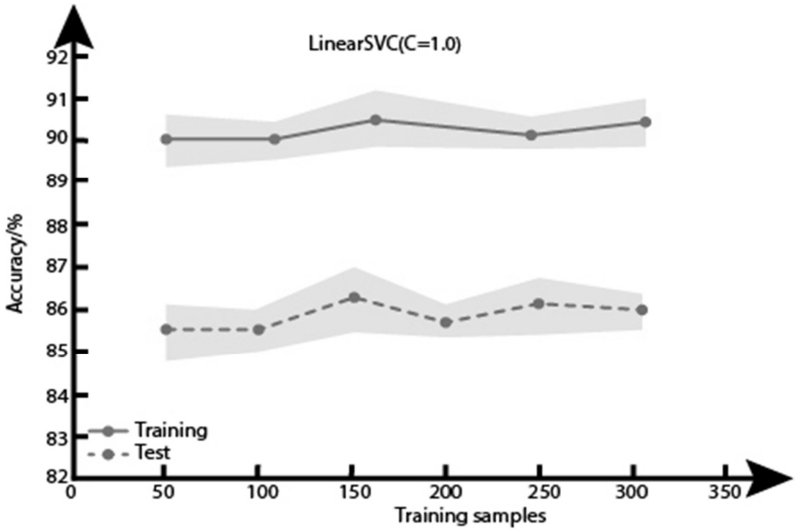


Figure 6.20 The over-fitting case when creating a learning model using the linear-SVC algorithm with a small data set up to 160 samples

Enlarege the Data Set Size To Cover More Data Points and Feature Patterns

- Enalrge the training data set will make them more repersentative to catch the variety, veracity and volume features
- The noise effects will be reduced to results in high biases
- Manual labeling is added to separate some artificial sample space.
- Sample data set can be also transformed to balance the distribution in different feature dimensions

Under-Fitting Models

- This is the case when the model produced by a given training set ends up a very low score performance, which is far below the user's expectation.
- Under-fitting phenomenon implies that a poor training set was chosen.
- The trained model cannot perform well at all on real testing data sets.
- Therefore, the model so obtained is totally unacceptable to users. .

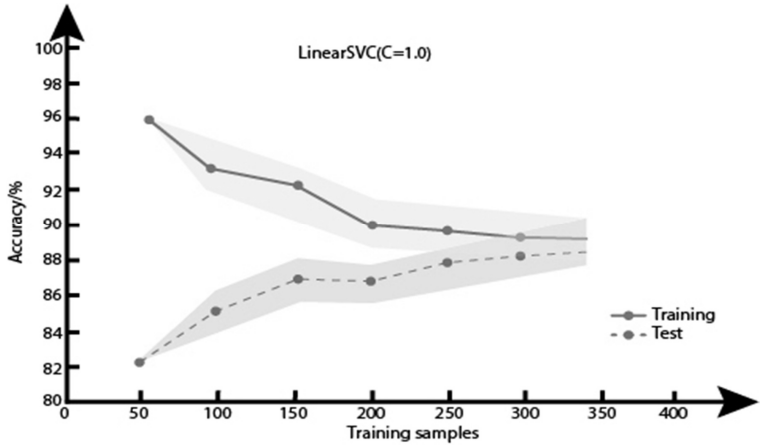


Figure 6.21 Reducing the model overfitting effects by enlarging the training set to 350 samples

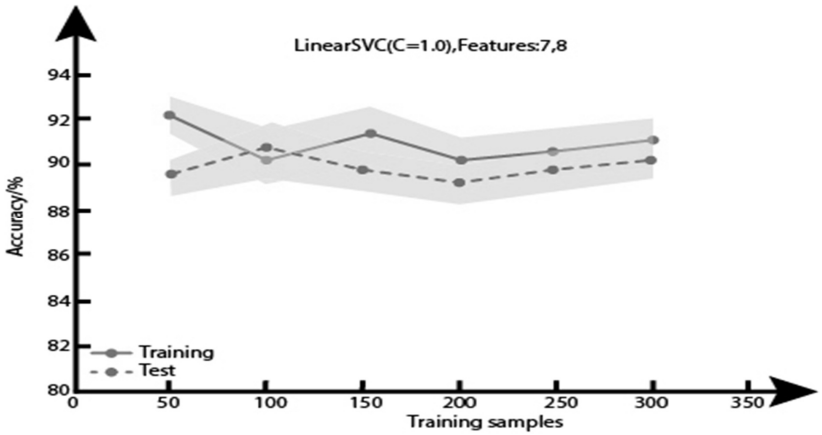
Feature Screening and Dimension Reduction (1)

- Revealing the correlation between features, one can cut off some features.
- Such feature screeninbg may reduce either the overfitting or even the under-fitting effects, depending on feature distributions
- Those features with limited repersentations in the data pace could be eliminated.

Feature Screening and Dimension Reduction (2)

- The method is known as feature screening or dimension reduction
- Association analysis or correlation analysis may help eliminate some weak dimensions.
- Principal Component Analysis goes the extreme to concentrate on the key features based on a spectrum analysis of the harmonics.

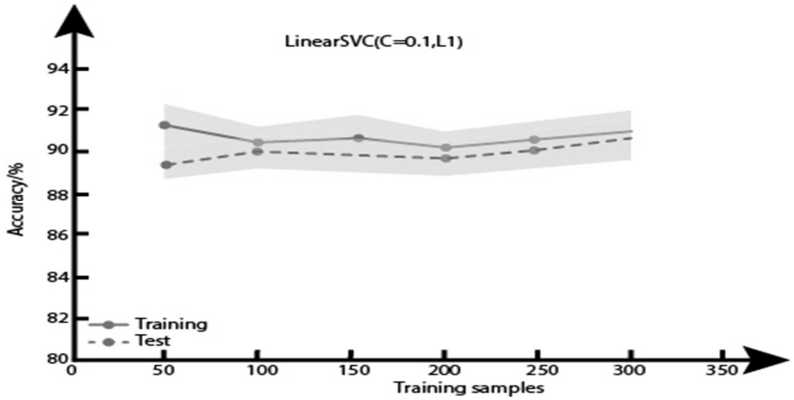
Moving Up the Testing Scores by Enlarging the Training Samples



Methods To Avoid Model Under-Fitting Effects

- Under-fitting is often resulted from poor samples collection.
- Miss Match between the machine learning algorithm and problem domain environment
- Consider the case of SVM (Simple Vector Machine) model for solving a classification problem. If the problem is not linearly speratable , one must switch to the nonlinear SVM model

Underfitting due to Linear–SVC Algorithm Used.



Machine Learning Model Selection Options: (1)

Loss functions express the discrepancy between the trained model prediction and the actual problem instances. The loss function reveals the effects of losing the expected performance of an ML algorithm. We consider below 5 loss functions.

.

- **Zero-one loss function:** This policy offers a very sharp division between success and failure. The 0-1 loss function just counts the number of miss-predictions in the classification problems. It is not practical in real-life applications.
- **Hinge loss function:** This is often used in SVM (Support Vector Machines) applications for its relative strength to reflect the unusual sensitivity to noise effects. This function is not supported by the probabilistic distributions.

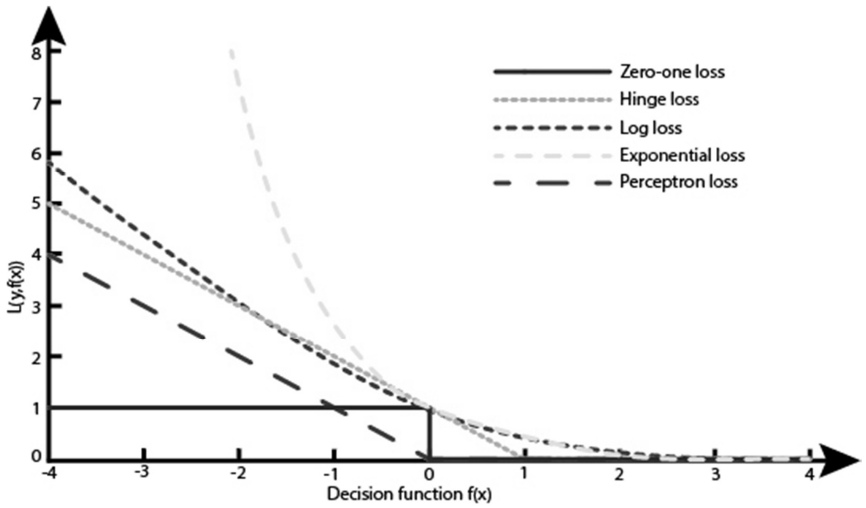


Figure 6.24: Effects of using different loss functions in machine learning model selection.

Machine Learning Model Selection Options: (2)

- **Log loss function: This reflects a probabilistic distribution.**
  - **The log-loss function is suitable for multi-classification problems.**
  - **The shortcoming lies in lower sensitivity to noises.**

- **Exponential loss function: This has been applied in AdaBoost, very sensitive to crowd and noises, but effective to deal with boosting algorithms.**

Machine Learning Model Selection Options: (3)

- **Perceptron loss function:**
  - **This is a variation from the hinge loss, which imposes heavy penalty to miss judgement of the boundary points**
  - **The perceptron loss is satisfied with accurate classification from sample data.**
  - **The advantage is its simplicity. The shortcoming lies in the fact it offers a weaker for lack of max-margin boundary.**

Summary of Model Fitting :

- **Choose a good sample dataset which is sufficicent large and repersentative of typical data behavior**
- **Training the model to fit with the sampling data**
- **Crosse validation of the model with testing data**
- **Modify the training set with more repersentative data points or change some features**
- **Modify the Machine Learning algorithm such as dimension reduction, etc.**
- **Use alternating methods to optimize the performance such as use ensembles or extending trees to forest**

Popular Social Networks in 2016

Table 5.8
Top social networks based on global user population in 2016

Social Network	Active Users	Social Network	Active Users
Facebook	1.65 billion	Twitter	320 million
WhatsApp	1.00 billion	Baidu Tieba	300 million
QQ	853 million	Skype	300 million
WeChat	697 million	Viber	249 million
Qzone	640 million	Sina Weibo	222 million
Tumblr	555 million	Line	215 million
Instagram	400 million	Snapchat	200 million

Four Top Social Networks by Population

Table 5.9
Summary of popular social networks and web services provided

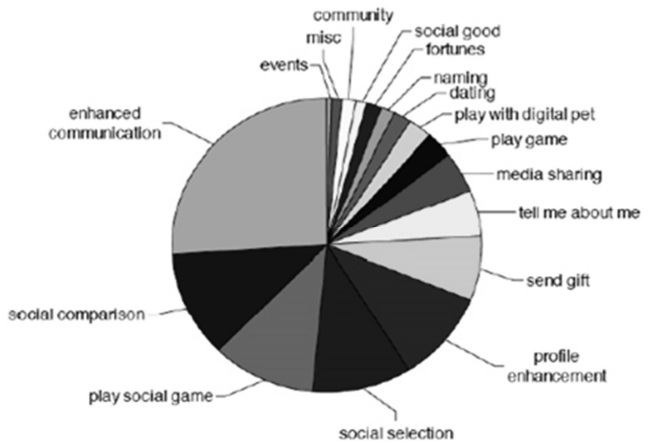
Social Network, Year, and Website	Registered Active Users	Major Services Provided
Facebook, *2004,* www.facebook.com	1.65 billion users, 2016	Content sharing, profiling, advertising, events, social comparison, communication, play social games, etc.
Tencent QQ *in China, 1999,* www.tencent.com	853 million users, 2016	An instant messaging service, online games, music, ebQQ, shopping, microblogging, movies, WeChat, QQ Player, etc.
Linkedin, *2002,* www.linkedin.com	364 million users, 2015	Professional services, online recruiting, job listings, group services, skills, publishing, advertising, etc.
Twitter, *2006,* www.twitter.com	320 million users, 2016	Microblogging, news, alerts, short messages, rankings, demographics, revenue sources, photo sharing, etc.

Social-Economic Impact of Social-Media Services

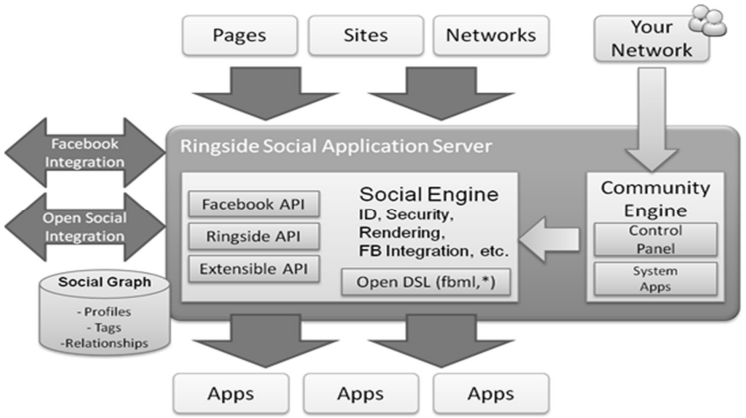
Table 5.7
Social media corporate functions weighted by social-economic impact

Corporate Function	Res. and Develop.	Marketing	Customer Service	Sales	Human Resources	Organization
Blogs	Low	Medium	Low			
Business Networks					Very high	Low
Collaborative Projects	Very high					Very high
Enterprise Networks	High				Medium	
Forums	Medium	Low	Very high		Low	
Microblogs		High				
Photo Sharing		Medium				
Products/Services Review	Low	Medium		Very high		
Social Bookmarking		Medium				
Social Gaming		Medium				
Social Networks	Low	Very high	Medium		Low	Low
Video Sharing		Very high	Low			
Virtual Worlds	Low	High		Low		

Facebook Applications



Facebook Web Site Architecture



Facebook Feature Capabilities

Function	Implementation
Profile Page	Combined Profile: profile picture, bio information, friends list, user's activity log, public message board, other components selectively displayed
Social Graph Traversal	Access through users' friends list on profile pages, with access control
Communication Tools	Internal Email-like Message: send and receive private messages among friends Instant Messaging: Accessed on the webpage, or through 3rd party client Public Message Board: "Wall", with access control Update Status: A short message, like micro-blogging, with access control
Share d Information	Photo Album: Built-in, with access control, Links: Post links to outside URL, will appear on the activity log, Videos: Embedded outside videos on profile page
Access Control	1. Every items of profile page can be set access control in four levels:, only me, 2. only friends, 3, friends of friends, 4. everyone
API	Games, calendars, mobile clients

Facebook Service Functionality

Table 5.10
Service functionality of the Facebook platform

Function	Short Description
Profile Pages	Profile picture, bio information, friends list, user's activity log, public messages
Graph Traversal	Access through user's friends list on profile page, with access control
Communication	Send and receive messages among friends, instant messaging, and microblogging
Shared Items	Photo album with built-in access control, embedded outside videos on profile page
Access Control	Access control levels: only me, only friends, friends of friends, and everyone
Special APIs	Games, calendars, mobile clients, etc.

C;loud Provider APIs for Social-Media Applications

Table 5.11
Social media application programming interfaces (APIs)

API Name	Functionality	Protocol Applied	Data Format	Security
Facebook Graph API	Facebook social graph processing, community detection, finding friends, etc.	REST	JSON	OAuth
Google+API	To provide access to Google+, a social media website with links, status, and photo options	REST	JSON	API key, OAuth
Social Mention API	Programmatic access to interact with Social Mention website, a RESTful API	HTTP	PHP	API key
Delicious API	Allows users to access, edit, and search for bookmarks	REST	JSON, RSS	OAuth, HTTP/ Basic
MySpace API	To access various MySpace functions and integrate application into MySpace	Javascript	Unknown	OAuth
Meetup API	To use the topics, groups, and events created by Meetup in their own applications	REST	JSON, XML KML, RSS	PAith, API key
FindMeOn API v.1,0	Programmatic access to the social media search and management functions of FindMeOn.	HTTP	JSON	API key
Cisco JTAPI	Cisco Java Telephony API allows Java applications to interact with Telephony resources	SOAP, HTTP	XML	SSL Support
YouTube Data API v3.0	Perform actions available on the YouTube website	REST, HTTP	JSON	API key

Social Network Analytics :

- Social Network Metrics
- Learning in Social Networks
- Relational Neighbor Classifier
- Relational Logistic Regression
- Collective Inferencing
- Egonets and Bigraphs

US Presidential Election Prediction in 2012

Figure 5.15
U.S. 2012 presidential election: Predicted 312 votes for Obama, which matches with the actual count. Source: Michael Cosentino (@Cosentino), Nov. 6, 2012, https://twitter.com/cosentino/status/266042007758200832

Social Media Cloud Processing (1)

Major challenges in exploring social media data over the cloud. Some may open up pportunities and some may reveal current IT shortcomings:

1. ***Unstructured social media data:*** **Traditional relational databases cannot support unstructured data, Thus we demand NoSQL processing of incomplete data from noisy and dirty sources [4]. These data are often short of veracity. Many blogs or social exchanges cannot be verified, requiring data filtering and integrity control.**

2. ***Social graphs, API, and visualization tools*** **are needed to handle unstructured social media data, effectively. This demands cost-effective clouds and distributed file systems to aggregate, store, process and analysis of big data. Bottom-up techniques are needed to uncover unknown structures and patterns.**

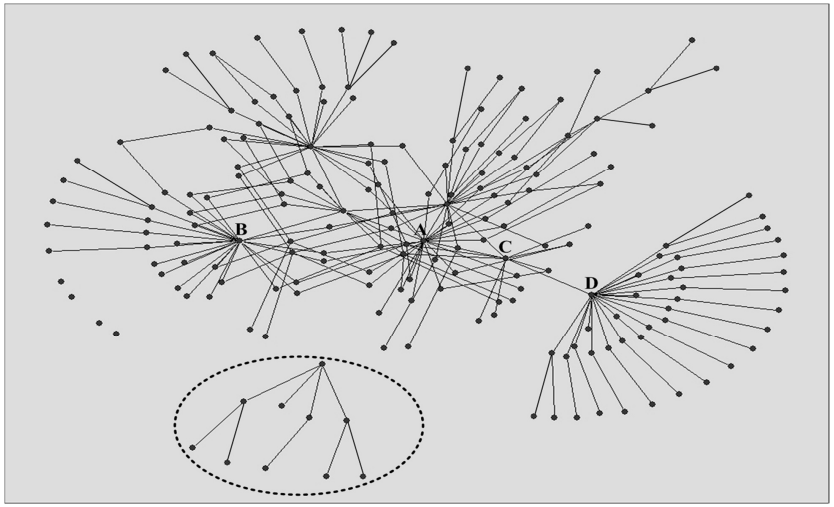
Social Media Cloud Processing (2)

Major challenges in exploring social media data over the cloud. Some may open up pportunities and some may reveal current IT shortcomings:

3. ***Machine learning and cloud analytics algorithms*** **are greatly on demand for supervised or unsupervised deep learning plus the use of feature extraction and pattern recognition techniques in predictive analytics. Data scientists must have sufficient domain knowledge, statistical data mining, social science, and programming skills from crossover domains to work cooperatively.**

4. ***Social media governance and security*** **demand data privacy, integrity control, SLA compliance, accountability, and trust management, etc. Social media security must be deployed on a global scale. Data privacy must preserved down to fine-grain data object or file levels.**

Graph Repersentation
of A Typical Social Network



Social Network Properties (1)

- ***Node degree :*** **The number of neighbors of a node.**
- ***Closeness*** **is the degree of an individual node being near to other nodes in a network directly or indirectly. It reflects the ability to access information through the network members.**
- ***Cohesion*** **is the degree to which actors (nodes) are connected directly to each other by cohesive bond.**
- ***Centrality and Centralization: Centrality indicates*** **the social power of a node based on how well they "connect" the network.**

Social Network Properties (2)

- ***Betweenness, Clos*****eness and *****Degr*****ee are all measures of centrality.**
- ***Centralized networks*** **have many of its links dispersed around one or a few nodes.**
- ***Decentralized network*** **is one in which there is little variation between the number of links each node possesses.**

Social Network Properties (3)

- ***Individual-level Density:*** **The degree a respondent's ties know one another or proportion of ties among an individual's nominees.**
- ***Network or global-level density*** **is the proportion of ties in a network relative to the total number possible ties.**
- ***Radiality*** **is the degree an individual's reachout into the network and provides influence.**

On-Line Social Networking Services

- Personal page or profiles for each user are linked by social connections.
- There is social graph traversal along specific social links or networks.
- Communication tools are shared between the participants or registered users.
- Special information like music, photos, videos, etc., is shared with friends or professional groups.
- Communities operate in special niche topic areas such as healthcare, sports, hobbies, etc.
- Customized software tools or databases are used to set up social network services.
- Strong customer loyalty creates viral membership growth.
- Social networks have revenues by selling premium memberships and access to premium content.

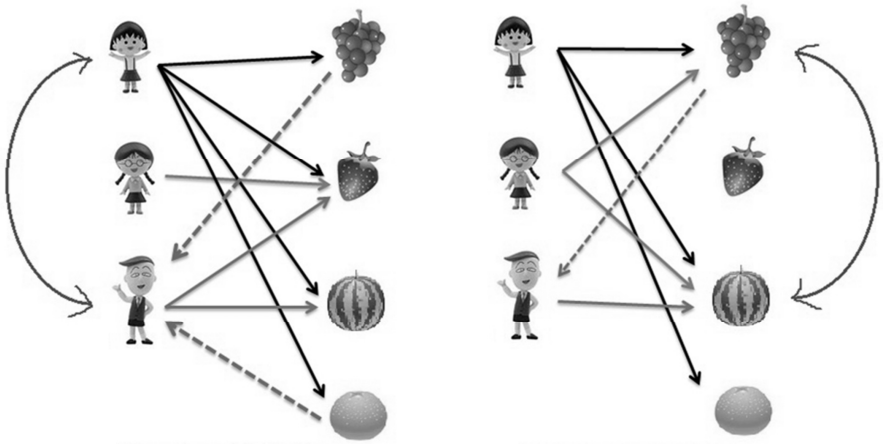
Social Network Properties (4)

- ***Structural cohesion*** **is the minimum number of embers who, if removed, would disconnect the group.**
- ***Structural equivalence*** **refers to the extent to which nodes have a common set of linkages to other nodes in the system. The nodes don't need to have any ties to each other to be structurally equivalent.**
- ***Structural hole :*** **Static holes can be filled by connecting one or more links to reach other points.**

Recommender Systems
in Movies, Tourism, Restaurants, etc.

- **Social or Collaborative Filtering: Polling the opinions of the mass to make decision based on ratings.**
- **Content-based Filtering: Recommend items based on features of products and ratings by other users.**
- **Demographic Filtering : Making decision based on demographic information of the user mass**

Recommender Systems
in Movies, Tourism, Restaurants and Many of our Daily-life Activities



Filtering Techniques for Recommender Systems

- **Knowledge-based Filtering : Making decision based on expertise knowledge and peer reputations, etc,**

- **Hybrid Filtering : Combine the advantages of some of the above filtering techniques to make smart decsions.**

Graph-Parallel Computation Methods in Sections 9.4.3 and 9.4.4

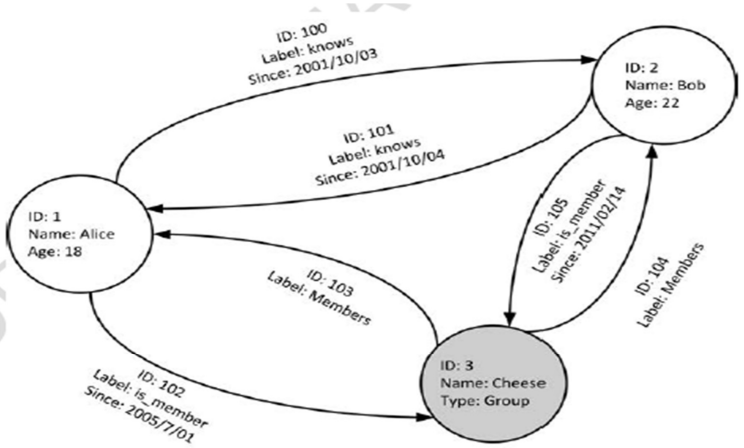


Figure 9.19
Graph analytics for revealing peer relationships in a social group. (Courtesy of Bill Schmarzo, "Graph Analytics 101," EMC2, January 2014.)

DIGITS 5 for Segmentation Workflow

For GPU-based Deep Learning , **Nvidia announced a new software platform, known as Nvidia DIGITS 5. This platform supports the training of neural networks in GPU-based deep learning applications. In many ways, this software package competes with TensorFlow in promoting distributed parallel execution on multiple GPU devices.**

.

The DIGIT 5 supports a database for photo image segmentation that enables the visualization of the output image from a segmented neural network. DIGITS model store is an open-source on-line knowledge base. It can download network description and pre-training models.

Example 9.7: Instance-Aware Image Segmentation in Neural Network Learning Operations

The *instance-aware image segmentation* (IAIS) refers to the segmentation of a given image into multiple segmemts or component parts. The neural network can learn the boundary of skeleton of each subdivided image segments. This very useful in practical imaging understanding applications. As shown in Fig.9.17, the IAIS system must be able to understand each class of image segments, even under the condition that some segment boundaries are fuzzy and not clearly distinguishable from surrounding image segments.

The left image has 5 persons lined up in a photo shot. The middle image is the segmented images of the five persons that are attached to each other without clear boundaries among them. The rightmost image is the same iamge after the IAIS treatment . Now their skeleton boundaries are plotted with edge extraction and labeling by different colors for different persons. The color labeling is one way to distinguish the individuals. This is the main concept of instace segmentation. The Facebook SharkMask has applied this technique to do image understanding and captioning. ■

Figure 9.18: A 5-person photo taken from the PASCAL VOC database. The segmented image is at the middle. The instance-segmented image is shown at the right for identification purpose. (Courtesy of www.nvidia.com, 2016).

Today's deep learning solutions rely almost exclusively on NVIDIA GPU-accelerated computing to train and speed up challenging applications such as image, handwriting, and voice identification. A deep learning system with Nvidia GPUs encourages parallel execution of image workloads. This can speed up networks by 10 to 75 times faster than using traditional CPUs. This hjas reduced the time of many image data training iterations from several weeks to just days. Nvidia has claimed that GPUs can achieve a 12 times in training *deep neural networks* (DNNs) over the use of CPU devices.

In general, the GPU approach offers faster AI application development. As a matter fact, today's computers can perform not only learning oerations, but also some kind of thinking in the image recognition process. . This opens up the opportunities in applications on robots, medicine, and self-driving cars. You can quickly design and deploy deep learning applications with real time responses. The GPUs are larged used in desktops, notebooks, servers and supercomputers around the world. Noe even GPU clouds have appeared in Amazon, IBM and Microsoft cloud platforms. The following example show a complete training solution in playing multiparty game. .

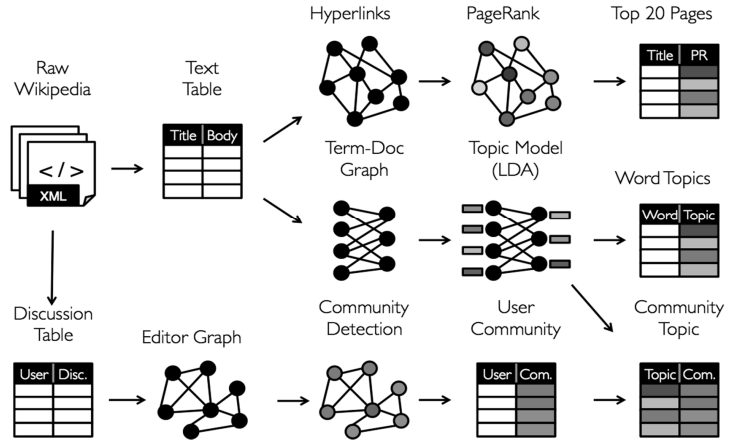


Figure 9.20: Converting raw data to tables and graphs berfore they are analyzed by graph algorithms. (Courtesy of Apache, https://www.spark.apache.org/docs/0.9.0/graphx-programming-guide, 2016 [2])

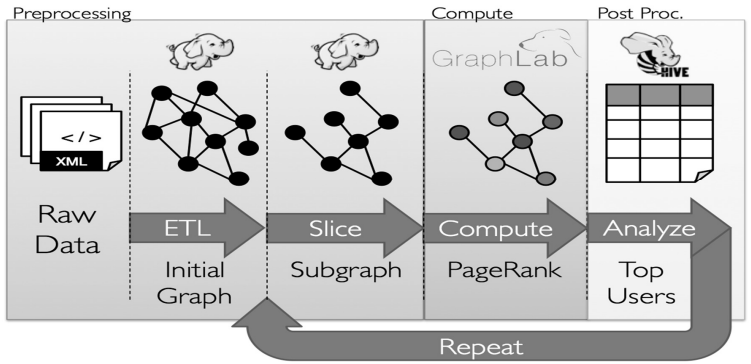


Figure 9.21: A typical graph analytics pipeline consisting of 5 functional stages. (Courtesy of Apache Spark, https://www.spark.apache.org/docs/0.9.0/graphx-programming-guide, 2016)

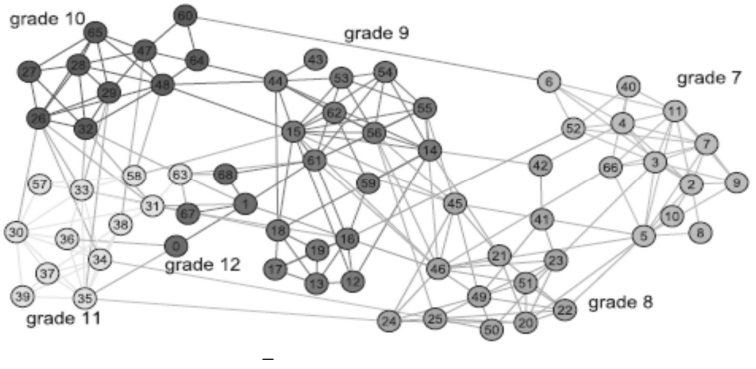


Figure 9.23 High scholl community formation based on grade class membership (Courtesy of Xie, J. et al, "Overlapping community detection in networks", *ACM Computing Survey*, August 2013 [15])